Adaptive Hierarchical Priors for High-Dimensional Vector Autoregressions[*]

Dimitris Korobilis
University of Essex[†]

Davide Pettenuzzo
Brandeis University[‡]

March 28, 2018

Abstract

This paper proposes a simulation-free estimation algorithm for vector autoregressions (VARs) that allows fast approximate calculation of marginal parameter posterior distributions. We apply the algorithm to derive analytical expressions for independent VAR priors that admit a hierarchical representation and which would typically require computationally intensive posterior simulation methods. The benefits of the new algorithm are explored using three quantitative exercises. First, a Monte Carlo experiment illustrates the accuracy and computational gains of the proposed estimation algorithm and priors. Second, a forecasting exercise involving VARs estimated on macroeconomic data demonstrates the ability of hierarchical shrinkage priors to find useful parsimonious representations. We also show how our approach can be used for structural analysis and that it can successfully replicate important features of news-driven business cycles predicted by a large-scale theoretical model.

Keywords: Bayesian VARs, Mixture prior, Large datasets, Macroeconomic forecasting

JEL Classifications: C11, C13, C32, C53

[*] We would like to thank Linda Bui, Andrea Carriero, Roberto Casarin, Sid Chib, Todd Clark, David Dunson, Sylvia Früwirth-Schnatter, Ana Galvao, Domenico Giannone, Jim Griffin, Maria Kalli, Gary Koop, Robert McCulloch, Ivan Petrella, Simon Price, Giorgio Primiceri, Francesco Ravazzolo, Giovanni Ricco, Barbara Rossi, Mark Steel, Allan Timmermann, and Frank Windmeijer. We would also like to acknowledge for their helpful comments participants at the University of Pennsylvania "Big Data in Predictive Dynamic Econometric Modeling" conference, the 23^{rd} Computing in Economics and Finance conference, the 2017 European Seminar on Bayesian Econometrics, the 2017 NBER Seminar on Bayesian Inference in Econometrics and Statistics, the 2017 Sheffield Macroeconomic Workshop, as well as seminar participants at the National Bank of Poland, the European Central Bank, Brandeis University, Brown University, and the Universities of Bristol, Kent, Warwick, and Southern California.

[†] Essex Business School, Wivenhoe Park, Colchester, CO4 3SQ, United Kingdom. d.korobilis@essex.ac.uk

[‡] Brandeis University, Sachar International Center, 415 South St, Waltham, MA. dpettenu@brandeis.edu

1 Introduction

There is ample evidence that exploiting large information sets can be beneficial for macroeconomic forecasting and structural analysis. While the early literature has established this fact in univariate applications (Stock and Watson, 2002), a more recent literature applies the same concept to multivariate vector autoregressions (VARs; see Banbura et al., 2010). Not surprisingly, a large body of this VAR literature relies on Bayesian methods, exploiting prior information as a way of achieving regularization and shrinkage. The early literature on vector autoregressions has focused on subjectively tuned priors such as the Minnesota prior (Doan et al., 1984; Litterman, 1979). In constrast, following advances in Bayesian computation, the current econometric literature highlights the importance of hierarchical priors as a way of eliciting the degree of prior informativeness objectively from the data.[1] Examples of this literature include Del Negro and Schorfheide (2004), who formulate a flexible procedure that allows the data to dictate how much weight should be attributed to prior moments coming from a general equilibrium model, and George et al. (2008), who propose a hierarchical prior that in a Gibbs sampler setting allows to search for VAR restrictions in an automatic and data-driven way.

Hierarchical shrinkage priors are structured using multiple layers of distributions, with upper level prior hyperparameters being conditioned on lower level hyperparameters. That way, very complex prior structures, such the famous Laplace prior that leads to the LASSO estimator (Tibshirani, 1996), can be decomposed into a series of tractable conditional prior distributions.[2] At the same time, hierarchical shrinkage can be seamlessly combined with independent priors, which have been shown to be important for VAR inference and forecasting. Notwithstanding their excellent properties and empirical successes, the vast majority of existing applications featuring hierarchical priors have been severely limited because of their reliance on computationally intensive Markov Chain Monte Carlo (MCMC) methods. For example, George et al. (2008) work with seven-variable models. In high-dimensions, when the VAR parameters proliferate at a polynomial rate, such simulation-based methods become

[1] In the statistics and machine learning literature, hierarchical priors are referred to as “sparse Bayesian learning” or “adaptive sparseness” priors, due to the fact that the informativeness of the prior is learned from the data; see Tipping (2001) and Figueiredo (2003).

[2] This feature, in turn, makes very natural in models with hierarchical priors the use of the Gibbs sampler, which is a technique for sampling from conditional posteriors.

computationally cumbersome, if not infeasible. A notable exception is Giannone et al. (2015) who, in order to estimate systems with more than 20 equations, rely on the natural conjugate prior to obtain posterior estimates for the degree of informativeness of their prior. However, their approach is restricted by the fact that the natural conjugate prior treats each VAR equation symmetrically, and imposes that the prior covariances of the coefficients in any two equations must be proportional to one another.[3]

In this paper we develop a new estimation algorithm for VARs under the proposed class of independent hierarchical shrinkage priors. Unlike the existing MCMC-based methods, the proposed approach is simulation-free and can be applied to models of very high dimensions. Also, unlike the approximation methods that rely on the restrictive natural conjugate prior (e.g., Banbura et al., 2010; Giannone et al., 2015), our suggested approach integrates hierarchical shrinkage within an independent prior setting. We capitalize on the efficient algorithm of van den Boom et al. (2015a) designed for a univariate regression, and further develop it to address the complexities of high-dimensional VARs. In particular, we first rewrite the VAR in its fully recursive form, which allows equation-by-equation estimation (Carriero et al., 2017). Next, as in van den Boom et al. (2015a), estimation relies on a simple transformation ("rotation") of each VAR equation which allows to approximate the joint posterior of the VAR coefficients as the product of a number of scalar marginal posterior distributions. The algorithm, therefore, breaks the multivariate estimation problem into a series of independent tasks each one involving a scalar parameter.[4] Finally, we extend and generalize the van den Boom et al. (2015a) algorithm, originally developed to implement variable selection, to three popular cases of hierarchical priors, namely (i) Normal-Jeffreys (Hobert and Casella, 1996), (ii) Spike-and-Slab (Mitchell and Beauchamp, 1988), and (iii) Normal-Gamma (Griffin and Brown, 2010), and show how to obtain analytical posteriors for the VAR coefficients and the elements of the VAR covariance matrix.

Using a Monte Carlo exercise, we find that our algorithm is as accurate as the comparable simulation-based methods but at a fraction of their computing time. At the same time, the

[3] This means that if we want to impose money neutrality in the VAR by shrinking to zero the coefficient of money in the equation for GDP, then the symmetry of the natural conjugate prior requires that the effect of money is removed from all other VAR equations in the system, even if money could still be a potentially useful predictor of, say, inflation.

[4] In the machine learning and graphical modeling literatures such procedures are known as *variable elimination*; see Barber (2012).

simulation-free nature of the algorithm means that there are no “convergence” or other similar numerical issues. Having established the numerical accuracy of our proposed algorithm, we then focus on two empirical exercises inspired by the recent literature on high-dimensional VARs. Our first application is a macroeconomic forecasting exercise using large-dimensional VARs of up to 124 equations. Banbura et al. (2010), Carriero et al. (2012), Carriero et al. (2017) and Koop et al. (2017) provide strong evidence that high-dimensional Bayesian VARs can consistently outperform smaller models. We show that when combined with the three hierarchical priors we focus on, our algorithm outperforms all competing methods in terms of forecast accuracy. Our second exercise involves using the new algorithm to estimate impulse response functions from an identified VAR. In particular, we simulate artificial time-series data from a large-scale DSGE model and show that our shrinkage methods can be used to obtain empirical VAR impulse response functions that follow closely the responses expected from the calibrated theoretical model.

The remainder of the paper is organized as follows. Section 2 describes in detail the estimation procedure we rely on to obtain analytical posteriors for the regression parameters in the presence of non-conjugate priors. Next, Section 3 examines the properties of three popular cases of hierarchical shrinkage priors and provides analytical derivations for the marginal posteriors of the coefficients of interest. Section 4 extends the methods described in Section 2 and Section 3 to the VAR case. After that, Section 5 describes the Monte Carlo exercise, while Section 6 is devoted to the macroeconomic forecasting application. Section 7 focuses on extending our algorithm to estimate impulse response functions using artificial data obtained from a large-scale DSGE model, and Section 8 offers some concluding remarks.

2 A new Bayesian estimation methodology

Before generalizing the estimation procedure to a VAR, our starting point is the following univariate regression model as in van den Boom et al. (2015a,b)

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{v}, \tag{1}$$

where $\boldsymbol{y} = (y_1, ..., y_T)'$ is a $T \times 1$ vector featuring our dependent variable, $\boldsymbol{X} = (\boldsymbol{X}_1', ..., \boldsymbol{X}_T')'$ is a $T \times k$ matrix involving T observations on k predetermined regressors, $\boldsymbol{\beta}$ is the corresponding

$k \times 1$ vector of regression coefficients, and $\boldsymbol{v} = (v_1, ..., v_T)' \sim \mathcal{N}\left(\mathbf{0}, \sigma^2 \boldsymbol{I}_T\right)$. When k is large, estimation of the high-dimensional posterior distribution $p\left(\boldsymbol{\beta}|\boldsymbol{y}\right)$ involves very costly operations (e.g. inversion of the high-dimensional matrix $\boldsymbol{X}$), and quickly becomes computationally demanding or even infeasible.

Following van den Boom et al. (2015a,b), we introduce an alternative approach to evaluate the marginal posteriors $\{p\left(\beta_j|\boldsymbol{y}\right)\}_{j=1}^{k}$ without the need to compute a number of high-dimensional integrals over the joint posterior distribution $p\left(\boldsymbol{\beta}|\boldsymbol{y}\right)$. We then proceed by approximating the full posterior $p\left(\boldsymbol{\beta}|\boldsymbol{y}\right)$ using the product of all k marginal posteriors.[5] Put simply, this approach works by transforming a complex and often intractable k-dimensional posterior evaluation problem into the product of k independent (and much simpler) estimation steps. We define the following rotation for each of the k columns in $\boldsymbol{X}$, one at a time

$$y_j^* = \boldsymbol{q}_j'\boldsymbol{y}, \qquad \widetilde{\boldsymbol{y}}_j = \boldsymbol{W}_j'\boldsymbol{y}, \tag{2}$$

where $j = 1, ..., k$, $\boldsymbol{q}_j = \boldsymbol{X}_j / \left\|\boldsymbol{X}_j\right\|$ is a $T \times 1$ unit vector in the direction of j-th column of $\boldsymbol{X}$ and $\boldsymbol{W}_j$ is an arbitrarily chosen $T \times T - 1$ matrix, subject to the constraint $\boldsymbol{W}_j\boldsymbol{W}_j' = \boldsymbol{I}_T - \boldsymbol{q}_j\boldsymbol{q}_j'$. Note that since the $T \times T$ orthogonal matrix $\boldsymbol{Q}_j = \left[\boldsymbol{q}_j, \boldsymbol{W}_j\right]$ is of full rank, the suggested rotation provides a one-to-one mapping between the original data $\boldsymbol{y}$ and the rotated data $\left(y_j^*, \widetilde{\boldsymbol{y}}_j'\right)'$. We show in Appendix A.1 that if we multiply both sides of (1) by $\boldsymbol{Q}_j$, after rearranging we obtain the following observationally equivalent regressions

$$\begin{aligned} y_j^* &= \left\|\boldsymbol{X}_j\right\| \beta_j + \boldsymbol{X}_{(-j)}^* \boldsymbol{\beta}_{(-j)} + v_j^*, \\ \widetilde{\boldsymbol{y}}_j &= \widetilde{\boldsymbol{X}}_{(-j)} \boldsymbol{\beta}_{(-j)} + \widetilde{\boldsymbol{v}}_j, \end{aligned} \tag{3}$$

where $\boldsymbol{X}_{(-j)}^* = \boldsymbol{q}_j'\boldsymbol{X}_{(-j)}$ is a $1 \times (k-1)$ vector, $v_j^* = \boldsymbol{q}_j'\boldsymbol{v}$ is a scalar, $\widetilde{\boldsymbol{X}}_{(-j)} = \boldsymbol{W}_j'\boldsymbol{X}_{(-j)}$ is a $(T-1) \times (k-1)$ matrix, $\widetilde{\boldsymbol{v}}_{\boldsymbol{j}} = \boldsymbol{W}_j'\boldsymbol{v}$ is a $(T-1) \times 1$ vector, and $\boldsymbol{X}_{(-j)} = \boldsymbol{X} \setminus \boldsymbol{X}_j$ denotes the $k-1$ columns of $\boldsymbol{X}$ after its j-th column has been removed. Similarly, $\boldsymbol{\beta}_{(-j)} = \boldsymbol{\beta} \setminus \beta_j$ denotes the $k-1$ elements of $\boldsymbol{\beta}$ after its j-th element has been removed. It also follows that the joint likelihood of the rotated data $\left(y_j^*, \widetilde{\boldsymbol{y}}_j'\right)'$ can be represented as

$$\left[\begin{array}{c} y_j^* \\ \widetilde{\boldsymbol{y}}_j \end{array}\right] \Bigg| \boldsymbol{\beta}, \sigma^2 \sim \mathcal{N}\left(\left[\begin{array}{c} \left\|\boldsymbol{X}_j\right\| \\ 0 \end{array}\right] \beta_j + \left[\begin{array}{c} \boldsymbol{X}_{(-j)}^* \\ \widetilde{\boldsymbol{X}}_{(-j)} \end{array}\right] \boldsymbol{\beta}_{(-j)}, \sigma^2 \boldsymbol{I}_T\right), \tag{4}$$

[5] This assumption implies posterior independence among coefficients, that is, $p\left(\boldsymbol{\beta}|\boldsymbol{y}\right) \equiv \prod_j p\left(\beta_j|\boldsymbol{y}\right)$. While such independence assumption can be very helpful for prediction, in Section 7 we also show how to modify this procedure in the context of a structural VAR in order to obtain the exact joint posterior.

where, due to the orthogonality of $\boldsymbol{Q}_j = \left[\boldsymbol{q}_j, \boldsymbol{W}_j\right]$, the variance of the rotated data is still σ^2. Most importantly, the rescaled regression in (3) separates the scalar y_j^*, which depends on β_j, from the remaining $T-1$ observations $\widetilde{\boldsymbol{y}}_j$, which are conditionally independent of the effect of β_j. At the same time, the form of the rescaled likelihood in (4) implies that y_j^* and $\widetilde{\boldsymbol{y}}_j$ do not share covariance terms, which ultimately means that we can treat (3) as two conditionally separable regression models. Combined, these last two equations provide insights on how to devise a simple two-step OLS procedure to estimate β_j. First, regress $\widetilde{\boldsymbol{y}}_j$ on $\widetilde{\boldsymbol{X}}_{(-j)}$ to obtain estimates for $\boldsymbol{\beta}_{(-j)}$ and σ^2, namely $\widehat{\boldsymbol{\beta}}_{(-j)}$ and $\widehat{\sigma}^2$. Next, condition on the regression variance $\widehat{\sigma}^2$ and regress $\left(y_j^* - \boldsymbol{X}^*_{(-j)}\widehat{\boldsymbol{\beta}}_{(-j)}\right)$ on $\|\boldsymbol{X}_j\|$ to obtain an estimate for β_j. Note that the estimates that we obtain from this two-step procedure are numerically identical to the OLS estimates we would recover if working with the original regression model in (1).[6]

We now exploit the form of the likelihood in (4), along with Bayes Theorem, to derive the following expression for the marginal posterior distribution $p\left(\beta_j|\boldsymbol{y}\right)$

$$\begin{aligned} p\left(\beta_j|\boldsymbol{y}\right) &= p\left(\beta_j|y_j^*, \widetilde{\boldsymbol{y}}_j\right) \\ &= \frac{p\left(\beta_j, y_j^*|\widetilde{\boldsymbol{y}}_j\right)}{p\left(y_j^*|\widetilde{\boldsymbol{y}}_j\right)} \\ &\propto p\left(y_j^*|\beta_j, \widetilde{\boldsymbol{y}}_j\right) p\left(\beta_j\right), \end{aligned} \tag{5}$$

where we have used the fact that $p\left(y_j^*|\widetilde{\boldsymbol{y}}_j\right)$ does not involve β_j, meaning it is simply a normalizing constant that can be removed, and also the result that $\widetilde{\boldsymbol{y}}_j$ does not convey any information about β_j, i.e. $p\left(\beta_j|\widetilde{\boldsymbol{y}}_j\right) \equiv p\left(\beta_j\right)$. Equation (5) shows that, thanks to the rotation in (2), the marginal posterior distribution of β_j is proportional to the rotated conditional likelihood $p\left(y_j^*|\beta_j, \widetilde{\boldsymbol{y}}_j\right)$ and the prior $p\left(\beta_j\right)$.[7] While we postpone our discussion on the prior distribution until the next section, it is of immediate interest to derive an expression for $p\left(y_j^*|\beta_j, \widetilde{\boldsymbol{y}}_j\right)$, and this is where we now turn our attention.

[6] This two-step approach is closely related to the traditional partitioned regression method (or "partial-time regression", using the terminology of Frisch and Waugh, 1933). There are however a number of important differences, which ultimately lead us to a procedure where we can estimate $\widehat{\boldsymbol{\beta}}_{(-j)}$ using $T-1$ observations, and estimate $\widehat{\beta}_j$ using a single observation. For additional details on the link with partitioned regression, see Appendix A.2.

[7] One implicit assumption we will rely on throughout is that the elements of $\boldsymbol{\beta}$ need to be a-priori independent, that is, $p\left(\boldsymbol{\beta}\right) = \prod_{j=1}^{k} p\left(\beta_j\right)$. This is a standard assumption in Bayesian analysis using hierarchical or other priors (e.g. Minnesota prior), since it is generally quite hard to objectively specify prior beliefs on the coefficients' cross-correlations.

Note that, from a Bayesian standpoint, the conditional likelihood function $p\left(y_j^*|\beta_j, \widetilde{\boldsymbol{y}}_j\right)$ can be interpreted as the predictive distribution of the "out-of-sample" data y_j^* given the "in-sample" data $\widetilde{\boldsymbol{y}}_j$, after the parameters $\boldsymbol{\beta}_{(-j)}$ and σ^2 have been integrated out. Using standard results for Bayesian predictive analysis (Koop, 2003), we show in Appendix A.3 that under a natural conjugate prior for $(\boldsymbol{\beta}_{(-j)}, \sigma^2)$ it follows that[8]

$$\begin{aligned} p\left(y_j^*|\beta_j, \widetilde{\boldsymbol{y}}_j\right) &= \left\|\boldsymbol{X}_j\right\| \beta_j + t_{2\overline{d}}\left(\overline{\mu}_j, \overline{\tau}_j^2\right) \\ &\approx \left\|\boldsymbol{X}_j\right\| \beta_j + \mathcal{N}\left(\overline{\mu}_j, \overline{\tau}_j^2\right), \end{aligned} \tag{6}$$

where

$$\overline{\mu}_j = \boldsymbol{X}_{(-j)}^* \overline{\boldsymbol{\beta}}_{(-j)}, \tag{7}$$

and

$$\overline{\tau}_j^2 = \frac{\overline{\psi}_{(-j)}}{\overline{d}} \left(1 + \boldsymbol{X}_{(-j)}^* \overline{\boldsymbol{V}}_{\boldsymbol{\beta}_{(-j)}} \boldsymbol{X}_{(-j)}^{*\prime}\right). \tag{8}$$

The exact formulas for the posterior moments $\overline{\boldsymbol{\beta}}_{(-j)}$, $\overline{\boldsymbol{V}}_{\boldsymbol{\beta}_{(-j)}}$, $\overline{\psi}_{(-j)}$, and $\overline{d}$ are standard to derive, and are also provided in Appendix A.3.

Two key remarks are in order. First, note that in equation (6) we have chosen to approximate a Student-t predictive distribution using a Normal distribution. An immediate question is how good an approximation this will be. Note that if σ^2 is known, then the formulas are exact. In other words, the rotated likelihood $p\left(y_j^*|\beta_j, \widetilde{\boldsymbol{y}}_j\right)$ is indeed Normal with the moments specified above. When σ^2 is unknown then the approximation can still be quite accurate, and the accuracy will increases with the sample size.[9] Second, equations (5) and (6) imply that it is now possible to compute the marginal posterior for β_j by solving a scalar linear regression model with normal data and known variance, $\overline{\tau}_j^2$. Most importantly, the fact that the variance of this regression is known and fixed means that we can derive analytically the marginal posterior for β_j even for priors that would normally require time-consuming simulation methods. This is a key result that we exploit in Section 3 to compute simulation-free marginal posteriors for a host of hierarchical shrinkage priors.

[8] While there are many alternative prior choices available for obtaining estimates of $(\boldsymbol{\beta}_{(-j)}, \sigma^2)$, we have chosen to rely on the natural conjugate prior because, among other things, it leads to proper posteriors for the regression parameters even when the number of parameters $(k-1)$ is larger than the total number of observations $(T-1)$, and at the same time leads to a closed-form expression for the conditional likelihood $p\left(y_j^*|\beta_j, \widetilde{\boldsymbol{y}}_j\right)$.

[9] This is related to the fact that a Student-t distribution with a sufficient number of degrees of freedom - typically 100 or more - converges to a Normal distribution.

The estimation steps resulting from the above analysis are summarized in Algorithm 1. While exact expressions depend on the choice of prior distribution, $p\left(\beta_j\right)$, here we give an example of how our algorithm would work with a generic prior.

Algorithm 1 Posterior estimation algorithm for a generic prior

```
for $j = 1$ to $k$
    STEP 1: PREPARE ROTATION MATRICES
        • Compute $\boldsymbol{q}_j = \boldsymbol{X}_j / \left\|\boldsymbol{X}_j\right\|$
        • Generate $\boldsymbol{W}_j$ from $\mathcal{N}\left(0,1\right)$ and apply QR decomposition to impose orthonormality
          of $\boldsymbol{Q}_j = \left[\boldsymbol{q}_j, \boldsymbol{W}_j\right]$
    STEP 2: APPLY ROTATION
        • Compute rotated data $y_j^*$ and $\widetilde{\boldsymbol{y}}_j$, $\boldsymbol{X}^*_{(-j)}$ and $\widetilde{\boldsymbol{X}}_{(-j)}$
    STEP 3: ESTIMATE AUXILIARY REGRESSION
        • Regress $\widetilde{\boldsymbol{y}}_j$ on $\widetilde{\boldsymbol{X}}_{(-j)}$, obtain moments of $p\left(\boldsymbol{\beta}_{(-j)} \middle| \sigma^2, \widetilde{\boldsymbol{y}}_j\right)$ and $p\left(\sigma^2 \middle| \widetilde{\boldsymbol{y}}_j\right)$ analytically
        • Derive moments of rotated likelihood, $\overline{\mu}_j$ and $\overline{\tau}_j^2$, analytically
    STEP 4: ESTIMATE PARAMETER OF INTEREST
        • Given $\overline{\mu}_j$ and $\overline{\tau}_j^2$, regress $\left(y_j^* - \overline{\mu}_j\right)$ on $\left\|\boldsymbol{X}_j\right\|$
        • Obtain moments of $p\left(\beta_j \middle| \boldsymbol{y}\right)$ analytically
end for
```

3 Hierarchical shrinkage priors

We now turn our focus to the prior for β_j $(j = 1, ..., k)$ in (5). While van den Boom et al. (2015a) focus on the problem of variable selection, we extend and generalize their approach to the following class of adaptive hierarchical priors for β_j,[10]

$$\begin{aligned} \beta_j \middle| \lambda_j^2 &\sim \mathcal{N}\left(0, \lambda_j^2 \underline{V}_{\beta_j}\right), \\ \lambda_j^2 &\sim G, \end{aligned} \tag{9}$$

where $\underline{V}_{\beta_j}$ denotes the part of the prior scale parameter chosen by the researcher, while λ_j^2 (or its square root, λ_j, depending on the specification) is a random variable with its own prior distribution, G.[11] Two observations are in order. First, the hierarchical form of the prior shows that conditional on the idiosyncratic scale parameter λ_j^2, the j-th regression coefficient β_j has a normal prior distribution. Combined with the approximation in (6), this is the key element that

[10] The assumption that the prior mean of β_j is zero is without loss of generality. All the results that follow can be trivially updated to allow for a non-zero prior mean.

[11] Alternatively, we could also refer to λ_j^2 as the local variance component. See for example Polson and Scott (2010).

will allow us to derive the posterior of β_j without resorting to simulation methods. Second, while the conditional prior for β_j is normal, the marginal prior of β_j, $p(\beta_j) = \int \mathcal{N}\left(0, \lambda_j^2 \underline{V}_{\beta_j}\right) dG\left(\lambda_j^2\right)$ ought not to be and, depending on the choice of G, can result in very different shapes, with possibly a large mass around zero and much heavier tails than a bell-shaped Normal prior, two features that will impose shrinkage in the regression model.

Within the class of adaptive hierarchical priors, we focus on three special cases for G, which in turn lead to three well-known Bayesian shrinkage estimators.

3.1 Normal-Jeffreys

The first choice of prior for λ_j^2 is a Jeffreys prior, i.e. $p\left(\lambda_j^2\right) \propto 1/\lambda_j^2$, which is fully uninformative about λ_j^2. Notice that this particular choice of prior for λ_j^2 leads to an improper marginal prior for β_j, i.e. $p(\beta_j) \propto |\beta_j|^{-1}$, a prior that is sharply peaked at zero and is similar to the popular Laplace prior, and therefore favors sparsity in the regression model (see for example Tipping, 2001; Figueiredo, 2003).

Thanks to the approximation in (6) and the conditional normality of the prior, it is straightforward to derive the marginal likelihood for y_j^* analytically. This takes the form

$$\begin{aligned} p\left(y_j^* \middle| \lambda_j^2, \widetilde{\boldsymbol{y}}_j\right) &= \int p\left(y_j^* \middle| \beta_j, \widetilde{\boldsymbol{y}}_j\right) p\left(\beta_j \middle| \lambda_j^2\right) d\beta_j \\ &= \mathcal{N}\left(y_j^* \middle| \overline{\mu}_j, \left\|\boldsymbol{X}_j\right\|^2 \lambda_j^2 \underline{V}_{\beta_j} + \overline{\tau}_j^2\right), \end{aligned} \tag{10}$$

where $\mathcal{N}(z|a,b)$ denotes the probability of a random variable z evaluated at a Normal distribution with mean a and variance b. Next, similar to the analysis of Giannone et al. (2015), we can choose the optimal shrinkage intensity λ_j^2 in (9) by maximizing (10), i.e.

$$\widehat{\lambda}_j^2 = \arg\max_{\lambda_j^2} \; p\left(y_j^* \mid \lambda_j^2, \widetilde{\boldsymbol{y}}_j\right). \tag{11}$$

We show in Appendix A.4 that the posterior estimate of λ_j^2 that maximizes the marginal likelihood takes the form

$$\widehat{\lambda}_j^2 = \max\left[0, \frac{\left(y_j^* - \overline{\mu}_j\right)^2 - \overline{\tau}_j^2}{\left\|\boldsymbol{X}_j\right\|^2 \underline{V}_{\beta_j}}\right]. \tag{12}$$

Finally, plugging the optimal shrinkage intensity $\widehat{\lambda}_j^2$ into (9) leads to the marginal posterior

$$p\left(\beta_j | \widehat{\lambda}_j^2, \boldsymbol{y}\right) \sim \mathcal{N}\left(\overline{\beta}_j, \overline{V}_{\beta_j}\right), \tag{13}$$

where both $\overline{\beta}_j$ and $\overline{V}_{\beta_j}$ depend on $\widehat{\lambda}_j^2$, and are given by

$$\overline{V}_{\beta_j} = \frac{\overline{\tau}_j^2 \widehat{\lambda}_j^2 \underline{V}_{\beta_j}}{\left\| \boldsymbol{X}_j \right\|^2 \widehat{\lambda}_j^2 \underline{V}_{\beta_j} + \overline{\tau}_j^2}, \qquad \overline{\beta}_j = \frac{\left\| \boldsymbol{X}_j \right\| \widehat{\lambda}_j^2 \underline{V}_{\beta_j} \left(y_j^* - \overline{\mu}_j \right)}{\left\| \boldsymbol{X}_j \right\|^2 \widehat{\lambda}_j^2 \underline{V}_{\beta_j} + \overline{\tau}_j^2}. \tag{14}$$

Notice, to conclude, that both the maximization in (12) and the prior moments in (14) only include scalar operations, so they are trivial to compute $\forall j \in [1, k]$.

3.2 Normal-Gamma

The second prior specification we consider within the class of hierarchical priors in (9) is the popular class of Normal-Gamma priors, studied in Griffin and Brown (2010) and extended to the VAR case by Huber and Feldkircher (2017). This prior assumes that $\lambda_j^2 \sim \mathcal{G}\left(\underline{c}_1, \underline{c}_2\right)$, where $\underline{c}_1$ and $\underline{c}_2$ denote the shape and scale of the Gamma distribution $\mathcal{G}$. To see the effect of the hyperparameters $\underline{c}_1$ and $\underline{c}_2$ on the shape of the marginal prior for β_j, the bottom panels of Figure 1 plot the marginal distribution of β_j for two different choices of $\underline{c}_1$ and $\underline{c}_2$. As a benchmark to compare against, the top left panel of the figure plots the empirical distribution of the non-hierarchical version of (9), where $\lambda_j^2 = 1$ is non-stochastic and $\underline{V}_{\beta_j} = 10.$[12] The bottom left panel plots the marginal prior of β_j when G is the Gamma density and the hyperparameters are set to $\underline{c}_1 = 1$ and $\underline{c}_2 = 2$. As it can be seen from this panel, this choice of hyperparameters generates a marginal prior for β_j that, compared to the benchmark bell-shaped Normal prior in the top left panel of the figure, shrinks towards zero at a much faster rate. Next, the bottom right panel of the figure considers the case where $\underline{c}_1 = 0.1, \underline{c}_2 = 2$. This choice leads to a much more intense shrinkage, with a clear spike around zero and tails that are significantly heavier than a Normal density.[13]

We can proceed in an analogous manner as in the Normal-Jeffreys case, and choose the

[12] For a large prior variance this can be considered a locally uninformative prior, while for small values of $\underline{V}_{\beta_j}$ it results in the ridge estimator.

[13] Notice that the Normal-Jeffreys prior is not plotted in this figure because it is an improper prior for λ_j^2, and leads to a marginal prior for β_j that does not integrate to one (and, thus, cannot be represented graphically). However, following Tipping (2001) we can think of the Normal-Jeffreys prior as a special case of a Normal-Inverse Gamma (IG) mixture, with $\lambda_j^2 \sim \mathcal{IG}\left(\underline{\alpha}_1, \underline{\alpha}_2\right)$ where $\underline{\alpha}_1, \underline{\alpha}_2 \to 0$. The Normal-IG mixture is the typical representation of the Student-t distribution, which is more peaked at zero compared to the Normal distribution. Therefore, the shrinkage induced by a Normal-Jeffreys can be broadly thought of as the limit of a Student-t prior with very large (infinite in practice) variance.

optimal shrinkage intensity by maximizing the posterior of λ_j^2,

$$\widehat{\lambda}_j^2 = \arg\max_{\lambda_j^2} \; p\left(y_j^* \mid \lambda_j^2, \widetilde{\boldsymbol{y}}_j\right) p\left(\lambda_j^2\right), \tag{15}$$

which, after taking logs, leads to the following maximization

$$\widehat{\lambda}_j^2 = \arg\max_{\lambda_j^2} \left\{ -\frac{1}{2} \ln\left(\overline{\tau}_j^2 + \|\boldsymbol{X}_j\|^2 \lambda_j^2 \underline{V}_{\beta_j}\right) - \frac{1}{2} \frac{\left(y_j^* - \overline{\mu}_j\right)^2}{\overline{\tau}_j^2 + \|\boldsymbol{X}_j\|^2 \lambda_j^2 \underline{V}_{\beta_j}} + \left(\underline{c}_1 - 1\right) \ln \lambda_j^2 - \underline{c}_2 \lambda_j^2 \right\}. \tag{16}$$

Once again, this is a straightforward maximization over scalar quantities, hence trivial to compute. Finally, plugging the optimal shrinkage intensity $\widehat{\lambda}_j^2$ into (9) leads to a marginal posterior for β_j with moments as in (14).

3.3 Spike-and-Slab

The third specification we consider for our hierarchical prior is the popular Spike-and-Slab prior, and follows very closely the approach described in van den Boom et al. (2015a). While it is possible to cast this prior in the hierarchical form of (9) (see for example Griffin and Brown, 2010, p. 175), we follow the literature and write this prior as an explicit mixture of distributions

$$\begin{aligned} \beta_j | \lambda_j &\sim \left(1 - \lambda_j\right) \delta_0 + \lambda_j \mathcal{N}\left(0, \underline{V}_{\beta_j}\right), \\ \lambda_j &\sim \mathit{Bernoulli}\left(\underline{\pi}_0\right), \end{aligned} \tag{17}$$

where δ_0 is the Dirac delta function at zero, while λ_j is now a Bernoulli random variable with mean $\underline{\pi}_0$ which, in turn, denotes the prior proportion of non-zero regressors in the model. As noted by Griffin and Brown (2010), the Spike-and-Slab and Normal-Gamma priors can lead to very similar forms of shrinkage. It is in fact possible to elicit the prior hyperparameters $\underline{c}_1$ and $\underline{c}_2$ of the Normal-Gamma prior and the prior inclusion probability $\underline{\pi}_0$ of the Spike-and-Slab prior in a way to similarly constrain most of the variation in the priors to a small set of regressors. Figure 1 makes this point explicitly, where in the top right panel we show the marginal prior of β_j for the Spike-and-Slab case and $\underline{\pi}_0 = 0.5$ (as with the other three panels, we set $\underline{V}_{\beta_j} = 10$). As it can be seen, the Spike-and-Slab prior with $\underline{\pi}_0 = 0.5$ leads to a marginal prior for β_j that behaves very much like the Normal-Gamma case when $\underline{c}_1 = 0.1$ and $\underline{c}_2 = 2$ (bottom right panel), placing a considerable mass at zero and featuring very heavy tails.

It follows that the posterior of λ_j is of the same form, that is $\lambda_j| \boldsymbol{y} \sim Bernoulli\left(\widehat{\pi}_j\right)$, where

$$\widehat{\pi}_j = p\left(\lambda_j = 1 | \, \boldsymbol{y}\right) = \frac{p\left(y_j^* \middle| \; \lambda_j = 1, \widetilde{\boldsymbol{y}}_j\right) p\left(\lambda_j = 1\right)}{p\left(y_j^* \middle| \, \lambda_j = 0, \widetilde{\boldsymbol{y}}_j\right) p\left(\lambda_j = 0\right) + p\left(y_j^* \middle| \, \lambda_j = 1, \widetilde{\boldsymbol{y}}_j\right) p\left(\lambda_j = 1\right)} \tag{18}$$

where $\widehat{\pi}_j$ is the posterior probability of inclusion (PIP) of predictor j in the regression model (not to be confused with a "p-value" or "significance level"). We show in Appendix A.5 that $\widehat{\pi}_j$ simplifies to

$$\widehat{\pi}_j = \frac{\mathcal{N}\left(y_j^* \middle| \overline{\mu}_j, \overline{\tau}_j^2 + \left\|\boldsymbol{X}_j\right\|^2 \underline{V}_{\beta_j}\right) \underline{\pi}_0}{\mathcal{N}\left(y_j^* \middle| \overline{\mu}_j, \overline{\tau}_j^2\right)\left(1 - \underline{\pi}_0\right) + \mathcal{N}\left(y_j^* \middle| \overline{\mu}_j, \overline{\tau}_j^2 + \left\|\boldsymbol{X}_j\right\|^2 \underline{V}_{\beta_j}\right) \underline{\pi}_0} \tag{19}$$

Finally, in this case the marginal posterior of β_j is equal to

$$\begin{aligned} p\left(\beta_j|\boldsymbol{y}\right) &= \int p\left(\beta_j|\lambda_j, \boldsymbol{y}\right) p\left(\lambda_j|\boldsymbol{y}\right) d\lambda_j \\ &= p\left(\lambda_j = 0|\boldsymbol{y}\right) p\left(\beta_j|\lambda_j = 0, \boldsymbol{y}\right) + p\left(\lambda_j = 1|\boldsymbol{y}\right) p\left(\beta_j|\lambda_j = 1, \boldsymbol{y}\right) \\ &= (1 - \widehat{\pi}_j)\delta_0 + \widehat{\pi}_j \mathcal{N}\left(\overline{\beta}_j, \overline{V}_{\beta_j}\right) \end{aligned} \tag{20}$$

where $\overline{\beta}_j$ and $\overline{V}_{\beta_j}$ are again given by (14) in the special case when $\widehat{\lambda}_j = 1$.

4 Application to BVAR estimation

Up to this point, we have focused our exposition on a univariate regression model. We now extend the current setup to a dynamic, multivariate setting, with a particular focus on the problem of estimating and forecasting with large-dimensional VARs. Consider the following n-dimensional VAR(p) model,

$$\boldsymbol{y}_t = \boldsymbol{c} + \boldsymbol{A}_1 \boldsymbol{y}_{t-1} + \ldots + \boldsymbol{A}_p \boldsymbol{y}_{t-p} + \boldsymbol{\varepsilon}_t, \quad t = 1, ..., T, \tag{21}$$

where $\boldsymbol{y}_t$ is an $n \times 1$ vector of time series of interest, $\boldsymbol{c}$ is an $n \times 1$ vector of intercepts, $\boldsymbol{A}_1, ..., \boldsymbol{A}_p$ are $n \times n$ matrices of coefficients on the lagged dependent variables, and $\boldsymbol{\varepsilon}_t \sim \mathcal{N}\left(\boldsymbol{0}, \boldsymbol{\Omega}\right)$, with $\boldsymbol{\Omega}$ an $n \times n$ covariance matrix. We next rewrite the original VAR model in (21) in a recursive form, which allows to estimate the VAR coefficients $\{\boldsymbol{c}, \boldsymbol{a}\}$ and the elements of the covariance matrix $\boldsymbol{\Omega}$ one equation at a time. This, in turns, allows us to readily apply the estimation method we presented in Section 2 to the VAR, by iterating recursively through a collection of univariate regressions.[14]

[14] Following standard results in multivariate models, one can factorize the covariance matrix $\boldsymbol{\Omega}$ into a diagonal matrix of variance terms and a lower triangular matrix of covariance terms. This factorization allows the covariance

From a computational perspective there are at least two ways one can re-write the reduced-form VAR in (21) as a recursive system. For example, Koop et al. (2017) rely on a recursive structural VAR representation. Here we use an alternative recursive form that is due to Carriero et al. (2017). We begin by decomposing the VAR covariance matrix $\boldsymbol{\Omega}$ in (21) as $\boldsymbol{\Omega} = \boldsymbol{\Gamma}^{-1}\boldsymbol{\Sigma}\left(\boldsymbol{\Gamma}^{-1}\right)'$, where

$$\boldsymbol{\Gamma}^{-1} = \begin{bmatrix} 1 & 0 & ... & 0 & 0 \\ \gamma_{2,1} & 1 & \ddots & \vdots & \vdots \\ \vdots & \ddots & \ddots & 0 & 0 \\ \gamma_{n-1,1} & ... & \gamma_{n-1,n-2} & 1 & 0 \\ \gamma_{n,1} & ... & \gamma_{n,n-2} & \gamma_{n,n-1} & 1 \end{bmatrix}, \tag{22}$$

and $\boldsymbol{\Sigma} = diag\left(\sigma_1^2, ..., \sigma_n^2\right)$. Thanks to this decomposition, it becomes possible to rewrite the i-th equation of the VAR $(i = 1, ..., n)$ as[15]

$$y_{i,t} = c_i + \boldsymbol{a}_{i,\cdot}\boldsymbol{Z}_t + \gamma_{i,1}\sigma_1 u_{1,t} + ... + \gamma_{i,i-1}\sigma_{i-1}u_{i-1,t} + \sigma_i u_{i,t}, \tag{23}$$

where c_i is the scalar intercept, $\boldsymbol{Z}_t = \left[\boldsymbol{y}_{t-1}', ..., \boldsymbol{y}_{t-p}'\right]'$ is a $np \times 1$ vector containing all p lags of $\boldsymbol{y}_t$, $\boldsymbol{a}_{i,\cdot} = [a_{i,1}, ..., a_{i,np}]$ denotes the corresponding vector of coefficients, $u_{1,t}, ..., u_{i-1,t}$ and $\sigma_1, ..., \sigma_{i-1}$ are the VAR structural residuals and standard deviations from all the previous $i-1$ equations, and $\gamma_{i,1}, ..., \gamma_{i,i-1}$ their associated coefficients. Next, let $\boldsymbol{X}_{i,t} = (\boldsymbol{Z}_t', \sigma_1 u_{1,t}, ..., \sigma_{i-1}u_{i-1,t})$ and rewrite (23) as

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta}_i + \boldsymbol{v}_i, \tag{24}$$

where $\boldsymbol{y}_i = (y_{i,t}, ..., y_{i,T})'$, $\boldsymbol{X}_i = \left(\boldsymbol{X}_{i,1}', ..., \boldsymbol{X}_{i,T}'\right)'$, $\boldsymbol{\beta}_i = (c_i, \boldsymbol{a}_{i,\cdot}, \gamma_{i,1}, ..., \gamma_{i,i-1})'$, and $\boldsymbol{v}_i = (\sigma_i u_{i,1}, ..., \sigma_i u_{i,T})'$. With the i-th equation of the VAR now in the same form as (1), we can straightforwardly apply the algorithm in Section 2 to the VAR, one equation at a time.

In particular, we will focus here on the generic β_{ij}, the j-th element of the vector $\boldsymbol{\beta}_i$ ($j = 1, ..., k_i$, while $k_i = np + i$ denotes the total number of coefficients in the i-th equation of the VAR). As in Section 2 we rely on the natural conjugate prior for $(\boldsymbol{\beta}_{(i,-j)}, \sigma_i^2)$ and follow the approach described in equations (6)-(8) and Appendix A.3 to integrate them out and obtain equation i's rotated likelihood, $p\left(y_{ij}^* \middle| \beta_{ij}, \widetilde{\boldsymbol{y}}_{ij}\right)$. However, instead of combining the use of the

terms to be treated as contemporaneous right-hand side predictors in each equation of the VAR and, because of the imposed recursive ordering, allows to estimate the VAR equation-by-equation; see Hausman (1983) for an early discussion of this approach.

[15] We provide additional details in Appendix A.6.

natural conjugate prior with random projection methods as in van den Boom et al. (2015a,b), we build on the successful approach of Banbura et al. (2010) and integrate out the effects of these parameters using a natural conjugate prior with Minnesota-type moments.[16] Similarly, we can modify the hierarchical prior in (9) to work with the VAR i-th equation by re-writing it as follows

$$\begin{aligned} \beta_{ij}|\,\lambda_{ij}^2 &\sim \mathcal{N}\left(0, \lambda_{ij}^2\underline{V}_{\beta_{ij}}\right), \\ \lambda_{ij}^2 &\sim G. \end{aligned} \tag{25}$$

One final point worth mentioning is that an added benefit of the procedure in (24)-(25) is that we can now apply our hierarchical shrinkage priors also to the coefficients $\gamma_{i,1}, ..., \gamma_{i,i-1}$, thus, explicitly providing shrinkage to the contemporaneous covariance elements in the VAR.

The outcome of this procedure is a flexible estimation method that provides closed-form posterior inference for VAR parameters in high-dimensional settings. Compared to the existing approaches in the literature, our method has the added benefit that it can work with both independent and hierarchical priors while at the same time requiring very minimal prior tuning, in this way allowing for individualized shrinkage on each VAR coefficient in a computationally very efficient way. In contrast, the competing approaches that provide closed-form solutions for the parameter posteriors either rely on the very restrictive natural conjugate prior (Banbura et al., 2010; Giannone et al., 2015) or make some other strong assumptions.[17] However, while our procedure is both flexible and computationally efficient it is an approximation method, and it imposes that all the elements of the vector $\boldsymbol{\beta}$ are a-posteriori uncorrelated. In the next two sections, we will devote significant space to showing how this approximation does not harm the forecasting performance of our method, and that the proposed procedure is at least as accurate as its MCMC-based hierarchical counterparts but at a fraction of their computing time. Furthermore, the a-posteriori independence leads to further computational benefits when using Monte Carlo integration to calculate functions of the VAR coefficients, such as for example

[16] Following Banbura et al. (2010), we work with $\boldsymbol{\beta}_{(i,-j)}|\sigma_i^2 \sim \mathcal{N}\left(\underline{\boldsymbol{\beta}}_{(i,-j)}, \sigma_i^2\lambda_{(i,-j)}\underline{\boldsymbol{V}}_{\boldsymbol{\beta}_{(i,-j)}}\right)$, $\sigma_i^2 \sim \mathcal{IG}\left(\underline{\psi}_i, \underline{d}_i\right)$, and set the overall shrinkage intensity on the elements of $\boldsymbol{\beta}_{(i,-j)}$ to $\lambda_{(i,-j)} = 0.1$. We also considered as a robustness check the possibility of optimizing $\lambda_{(i,-j)}$ as in Banbura et al. (2010), but found this to have no real effects on our results. As for the prior on σ_i^2, we opted for a non-informative prior, setting $\underline{\psi}_i = 0.01$ and $\underline{d}_i = 0.01$.

[17] For example, Litterman (1979) does estimate a VAR with independent priors but at the cost of fixing the covariance matrix to a first-step OLS estimate. Such an assumption underestimates parameter uncertainty in the covariance matrix and, as a by-product, in the predictive densities.

multi-step-ahead predictive densities. In this case, both MCMC and other analytical approaches (Banbura et al., 2010) requires to draw multiple times from the high-dimensional parameter posterior. In our case, the posterior independence leads to the possibility of breaking this sampling problem into $k_i \times n$ independent tasks, leading to further gains in computational efficiency.

5 Monte Carlo analysis

In this section we evaluate numerically the new approach using simulated data. The purpose of this exercise is manifold. First, we want to assess the numerical precision of the new estimation method. We have already argued that if we apply OLS (equivalently, a diffuse, objective prior) to the two-stage rotated regression in (3), we will obtain coefficients estimates that are identical to those we would obtain from OLS applied to the original regression problem in (1). However, it is important to evaluate whether the new estimation algorithm works well under a wide variety of Bayesian priors that will lead to biased penalized estimators. Second, we want to establish whether the three hierarchical priors introduced in Section 3 have good shrinkage properties when applied to a VAR setting and a finite amount of data. While the properties of such priors have been thoroughly examined and discussed in the literature, it is important to assess how the approximations we have introduced affect their performance. Finally, we want to obtain a measure of how well the proposed method fares against popular methods in recovering the true VAR coefficients.

5.1 Setup of Monte Carlo experiment

In order to investigate the importance of shrinkage as a function of the VAR size, we consider VARs of three different dimensions with $n = 3$, $n = 7$, and $n = 20$ endogenous variables. For each VAR dimension, we generate 1,000 datasets with $T = 150$ observations each. In all three cases, we set the number of lags to $p = 2$. The data generating process is that of a sparse VAR, where we allow the sparsity pattern to be random. We first model the persistence of each variable in the VAR by setting the first own lag coefficient to be in the range $[0.4, 0.6]$, i.e.

$$\boldsymbol{A}_1 = diag\left(\rho_1, \rho_2, ..., \rho_n\right), \tag{26}$$

where $\rho_i \sim \mathcal{U}(0.4, 0.6)$, $i = 1, ..., n$. The coefficients on the subsequent own lags, $(A_l)_{i,i}$ are then generated according to the rule that $(A_l)_{i,i} = (A_1)_{i,i}/l^2$ $(l = 2, ..., p)$, implying a geometric decay in their magnitudes, with the more distant lags having a lesser impact.[18] As for the coefficients on the other lags, we set them according to the following rule:

$$(A_l)_{i,j} = \begin{cases} \mathcal{N}\left(0, \sigma_A^2\right) & \text{with prob } \xi_A \\ 0 & \text{with prob } (1 - \xi_A) \end{cases} \qquad l = 1, ..., p, \quad i \neq j, \tag{27}$$

where $\xi_A \in (0, 1)$ is the probability of obtaining a non-zero coefficient. We set $\sigma_A^2 = 0.1$ and calibrate the inclusion probability according the the VAR size by setting $\xi_A = 1/(n-1)$. This means, for example, that in a seven-variable VAR only 1/6 of the coefficients are expected to be non-zero. Next, we decompose the covariance matrix $\mathbf{\Omega}$ as $\mathbf{\Omega} = \mathbf{\Phi}\mathbf{\Phi}'$ where

$$\mathbf{\Phi} = \begin{bmatrix} 1 & 0 & ... & 0 \\ \varphi_{2,1} & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ \varphi_{n,1} & \cdots & \varphi_{n,n-1} & 1 \end{bmatrix}, \tag{28}$$

and generate the element of $\mathbf{\Phi}$ according to the following rule

$$\varphi_{i,j} = \begin{cases} \mathcal{U}(0, 1) & \text{with prob } \xi_{\mathbf{\Phi}} \\ 0 & \text{with prob } 1 - \xi_{\mathbf{\Phi}} \end{cases} \qquad i > j. \tag{29}$$

where we set $\xi_{\mathbf{\Phi}} = 0.5$.

Along with our proposed algorithm and the three priors described in Section 3 (**Normal-Jeffreys**; **Normal-Gamma**; **Spike-and-Slab**), we consider the following three competing estimation methods: OLS (**VAR**); hierarchical Minnesota shrinkage as in Giannone et al. (2015) (**BVAR-GLP**); stochastic search for VAR restrictions algorithm of George et al. (2008) (**SSVS**). The BVAR-GLP approach relies on Minnesota-type moments, so due to the fact that the generated VARs are all stationary we set the prior mean on the first own lag coefficient to 0.9. For all the remaining coefficients, we set the prior mean to zero (see Kadiyala and Karlsson, 1997, for a discussion of these choices). For consistency, we use the same prior means in all the other Bayesian approaches, including ours (that is, we modify the hierarchical prior in (25) to allow for a non-zero mean, which we denote with $\underline{\beta}_{ij}$). The remaining settings

[18] The relatively low value of ρ_i and the decay in the own lag coefficients is done to guarantee that all variables in the VAR are stationary. In practice, in all cases we examine the roots of the generated VAR coefficients and discard all simulated DGPs producing non-stationarity variables.

for the BVAR-GLP algorithm are the default ones suggested by the authors.[19] As for the SSVS algorithm, we follow George et al. (2008) and set (using the authors' notation) the prior inclusion probabilities to $p_i = q_{ij} = 0.5$, and the prior variances to $R = R_j = I$, $\tau_0 = \kappa_0 = 0.1$ and $\tau_1 = \kappa_1 = 1$. As for the remaining details of our approach, we set the prior variance in (25) to $\underline{V}_{\beta_j} = 10$. Also, in the Spike-and-Slab case we set the prior inclusion probability for all predictors to $\underline{\pi}_0 = 0.5$, while in the Normal-Gamma case we set $\underline{c}_1 = 0.1$ and $\underline{c}_2 = 2$.[20]

5.2 Results

We begin by drawing attention to the estimated shrinkage intensity implied by our approach under the three different priors we considered. The top panels of Figure 2 and Figure 3 plot the empirical distribution of the average shrinkage intensity $\overline{\lambda}$ over the 1,000 Monte Carlo iterations for the three VAR sizes and for the Normal-Jeffreys and Normal-Gamma cases, respectively. In both figures, $\overline{\lambda} = \frac{1}{K}\sum_{i=1}^{n}\sum_{j=1}^{k_i}\widehat{\lambda}_{ij}$, where $K = \sum_{i=1}^{n} k_i$ denotes the total number of VAR coefficients, including the covariance terms in $\mathbf{\Phi}$. As one may expect, both in the case of the Normal-Jeffreys and the Normal-Gamma prior, the average shrinkage intensity becomes smaller as the VAR size increases, implying that more shrinkage is imposed in higher dimensions. This is a desirable feature of shrinkage estimation in VARs, and in line with previous findings in the literature; see Banbura et al. (2010) and their relevant discussion. This result is particularly clear in the case of the Normal-Gamma prior, where the empirical distribution of $\overline{\lambda}$ becomes more concentrated and informative as the VAR size increases.

A notable feature of our procedure is that it yields individualized shrinkage hyperparameters for each VAR coefficient, including the elements of the covariance matrix $\mathbf{\Phi}$. It would then be conceivable to expect that the VAR parameters which are equal to zero in the DGP should be accompanied by, on average, lower $\widehat{\lambda}_{ij}$'s. In order to verify this claim, the bottom panels of Figure 2 and Figure 3 plot the empirical distributions of the average shrinkage intensity $\overline{\lambda}$, after the individual $\widehat{\lambda}_{ij}$'s have been grouped according to whether the underlying VAR coefficients are equal to zero or not in the DGP. As expected, for both priors we find that the average

[19] Following Giannone et al. (2015), we specify the natural conjugate prior with the Minnesota moments by using a single shrinkage hyperparameter λ to control the overall tightness of the priors. We note however that in principle it would be possible to elicit a different prior hyperparameter for each equation in the model without foregoing the closed-form solution for the posteriors of all VAR parameters.

[20] In all cases, intercepts are left unrestricted using a diffuse prior. Note also that for both the SSVS algorithm and our estimation algorithms, we allow for shrinkage estimation of the sparse covariance terms $\varphi_{i,j}$.

shrinkage intensity of the zero VAR parameters (red bars) is significantly on the left of the average shrinkage intensity corresponding to non-zero VAR coefficients (blue bars). Notably, Figure 3 shows that for a large number Monte Carlo iterations, the average shrinkage intensity associated with the zero VAR coefficients is exactly zero, meaning that the hierarchical Gamma prior is capable of accurately flagging the irrelevant coefficients, shrinking all of them to zero. This result is more pronounced for the $n = 3$ and $n = 7$ VAR sizes, implying that for the larger $n = 20$ case, different values of the hyperparameters $\underline{c}_1, \underline{c}_2$ may be needed to achieve a similar result.

Figure 4 plots the distribution of the average posterior inclusion probabilities (PIPs) for the Spike-and-Slab prior, $\overline{\pi} = \frac{1}{K}\sum_{i=1}^{n}\sum_{j=1}^{k_i}\widehat{\pi}_{ij}$. In this case, due to the fact that there is a well-established alternative MCMC algorithm for VARs that relies on this prior, we contrast the results of our Spike-and-Slab hierarchical prior with those from the SSVS approach of George et al. (2008). In particular, the top panels of the figure plot the empirical distributions of $\overline{\pi}$ estimated with the SSVS algorithm, while the bottom panels plot the empirical distribution of $\overline{\pi}$ estimated using our algorithm and the Spike-and-Slab hierarchical prior. Once again, we separately plot the average PIPs corresponding to VAR parameters that are equal to zero (different from zero) in the DGP. As it can be seen from inspecting the figure, both algorithms are quite accurate at flagging which VAR parameters should be zero (or not), with the empirical distributions of the average PIPs from the zero VAR coefficients on the left of the corresponding non-zero coefficients' empirical distributions. Nevertheless, our algorithm performs visibly much better than the SSVS, with the estimated distributions being closer to zero and one (in the case of the SSVS algorithm, both distributions are close to 0.5 implying a decreased ability to determine if a VAR parameter is zero or not).

We next look at the effectiveness of the various methods in recovering the parameters of the true data generating process. To this end, for each of the approaches considered in this section, we compute the Mean Absolute Deviation (MAD), defined as

$$MAD^{(r,s)} = \frac{1}{K}\sum_{i=1}^{n}\sum_{j=1}^{k_i}\left|\beta_{ij}^{(r)} - \widehat{\beta}_{ij}^{(r,s)}\right|, \tag{30}$$

where s denotes the method used (VAR, BVAR-GLP, SSVS, Normal-Jeffreys, Normal-Gamma, Spike-and-Slab), $r = 1, ..., 1,000$ keeps track of the MC simulations, K denotes the total number of lag coefficients in the VAR, $\beta_{ij}^{(r)}$ is the true VAR coefficient from the r-th simulation, and $\widehat{\beta}_{ij}^{(r,s)}$ denotes the (posterior mean of the) corresponding estimate according to method s. Figure 5 shows the quartiles and median of the MAD statistic over all 1,000 Monte Carlo iterations, by means of box plots. For $n = 3$ the various shrinkage methods do not appear to improve much compared to OLS in recovering the true VAR parameters. However, as the VAR size increases, OLS begins to work less well. On the other hand, our estimation algorithm combined with the three hierarchical priors we introduced in Section 3 seems capable of accurately recovering the true VAR parameters, performing better than SSVS and as well or better than the BVAR-GLP method.

6 Macroeconomic forecasting

Combined with the simulation-free nature of our algorithm, the excellent properties of the hierarchical priors we introduced in Section 3 make them a very natural choice for a large dimensional VAR application. In this section, we investigate this claim empirically.

6.1 Data, models, and prior settings

We collect 124 quarterly variables for the US spanning the period 1959Q1 to 2015Q4.[21] The data, which are obtained from the Federal Reserve Economic Data (FRED) and are available at https://fred.stlouisfed.org, cover a wide range of key macroeconomic variables that applied economists monitor regularly, such as different measures of output, prices, interest and exchange rates, and stock market performance. We provide a full list of the data and their transformations in order to achieve stationarity in Appendix B. Out of the 124 series, we further distinguish seven "variables of interest", that is, key variables of interest which we will inspect very closely in order to evaluate how well the different models perform. These variables are: real gross domestic product (GDP), GDP deflator (GDPDEFL), and federal funds rate (FEDFUNDS), total employment (PAYEMS), unemployment rate (UNRATE), consumer prices (CPIAUCSL), and the 10-year rate on government securities (GS10).

We estimate VARs of three different sizes: Medium (the seven variables of interest plus an additional 13), Large (variables in medium plus an additional 20), and X-large (all 124

[21] For the variables which are originally observed at the monthly frequency, we transform them into quarterly series by computing the average of their monthly values within each quarter.

series available), that is, we consider 20, 40 and 124-variable VARs. All VARs have a lag length of $p = 5$. For each VAR size, we estimate a range of different models. In addition to the three hierarchical priors estimated using our simulation-free method, which we denote as N-J (Normal-Jeffreys), SNS (Spike-and-Slab), and N-G (Normal-Gamma), we consider six established methods for dealing with VARs of possibly large dimensions. The first three methods are based on the Minnesota prior with either optimal or pre-selected tuning of its shrinkage, one allows for Bayesian variable selection and model averaging, and two methods rely on factor shrinkage. In particular, we denote as BVAR-BGR the model of Banbura et al. (2010) who optimize the Minnesota shrinkage hyperparameter using a grid, while we denote as BVAR-GLP the model of Giannone et al. (2015) who introduce a hierarchical prior on the same Minnesota shrinkage hyperparameter and derive its posterior update formula.[22] The third method we consider is a BVAR with independent priors and Minnesota moments, which we denote as BVAR-IP. Compared to the previous two approaches, the BVAR-IP relies on non-conjugate priors and therefore requires the use of a Gibbs sampler. In order to guarantee large computational gains, we estimate this model using the algorithm of Carriero et al. (2017). Next, as a representative of simulation-based hierarchical shrinkage models, we consider the stochastic restrictions search algorithm of George et al. (2008), which we denote as SSVS. This algorithm is based on a mixture shrinkage prior, similar to the Spike-and-Slab prior we introduced in Section 3. Finally, we consider a dynamic factor model (denoted DFM), and a factor augmented VAR (denoted FAVAR); see Stock and Watson (2002) and Bernanke et al. (2005).

For the sake of comparability, whenever possible, we use the same exact prior settings. In particular, all Bayesian VAR models (including our three hierarchical prior and the SSVS method) feature the same Minnesota-based prior moments, which we write as

$$\underline{\beta}_{ij} = \begin{cases} 0.9 & \text{if own first lag} \\ 0 & \text{otherwise} \end{cases}, \qquad \underline{V}_{\beta_{ij}} = \begin{cases} \frac{1}{l_{ij}^2} & \text{if own lags} \\ \frac{\psi \times \widehat{\sigma}_i^2}{l_{ij}^2 \times \widehat{\sigma}_k^2} & \text{otherwise} \end{cases}, \tag{31}$$

where $i = 1, ..., n$, $j = 2, ..., np+1$, $\widehat{\sigma}_i^2$ ($\widehat{\sigma}_k^2$) is the OLS estimate of the variance of an AR(p) model on y_{it} (y_{kt}), $l_{ij} = \lfloor j/i \rfloor$ is the lag-length associated with the coefficient β_{ij} in the VAR, and ψ is

[22] Both the BVAR-BGR and BVAR-GLP approaches approximate inference using a natural conjugate prior which, as explained in the Introduction, has the disadvantage of symmetry across VAR equations, but the big advantage of leading to analytical expressions for the posterior moments of the VAR coefficients. Following the norm in the empirical literature, we implement both approaches by using a single shrinkage hyperparameter λ to control the overall tightness of the priors.

a hyperparameter that allows coefficients of variable k showing up in VAR equation i $(i \neq k)$ to shrink differently than own coefficients (k denotes the variable that the β_{ij} coefficient belongs to, i.e. $k = j - n\,(l_{ij} - 1)$).[23,24] Next, note that in our implementation of the BVAR-BGR, BVAR-GLP, and BVAR-IP models the shrinkage intensity is the same across all VAR coefficients i.e., using the notation in (25), $\lambda_{ij}^2 = \lambda^2$. In contrast, the SSVS prior of George et al. (2008) and our three hierarchical priors, N-J, SNS and N-G, do allow separate shrinkage intensities λ_{ij}^2. In particular, in the BVAR-BGR case we follow Banbura et al. (2010) and use a wide grid of possible λ^2 values. As for the BVAR-GLP case, the choice of the optimal shrinkage intensity is fully automatic.[25] Finally, when estimating the BVAR-IP model the overall prior tightness needs to be chosen a priori by the user, so we follow the recommendation of Sims and Zha (1998) and set $\lambda^2 = 0.2^2$. The other shrinkage hyperparameter ψ is set in all models to be a function of the VAR size, with $\psi = 0.001$ for the medium VAR, $\psi = 0.0001$ for the large VAR, and $\psi = 0.00001$ for the X-large VAR (note that the BVAR-BGR and BVAR-GLP models require $\psi = 1$). The remaining prior settings for the SSVS SNS, N-J, and N-G priors are: $\pi_0 = 0.1$, that is, our prior expectation is that only 10% of VAR coefficients are non-zero; $\underline{c}_1 = 0.1$, and $\underline{c}_1 = 2$. As for the prior hyperparameters specific to the SSVS we also set, using notation from George et al. (2008), $\tau_0 = \kappa_0 = 0.001$ and $\tau_1 = \kappa_1 = 10$. Finally, the DFM and FAVAR are estimated using principal components of the factors and a non-informative prior. We use the Bayesian information criterion (BIC) to select the optimal number of factors (minimum allowed is 1 and maximum is $\lfloor \sqrt{n} \rfloor$, with n the VAR size) and the optimal number of lags (ranging from one to five).

6.2 Measuring predictive accuracy

We use the first twenty five years of data, 1959:Q3–1984:Q4, to obtain initial parameter estimates for all models, which are then used to predict outcomes from 1985:Q1 $(h = 1)$ to 1985:Q4 $(h = 4)$. The next period, we include data for 1985:Q1 in the estimation sample, and use the resulting

[23] We denote with $\lfloor x \rfloor$ the floor of x, i.e. the largest integer less than or equal to x.

[24] Both the intercepts and the elements of $\mathbf{\Gamma}^{-1}$ are left unrestricted with flat and uninformative priors, i.e. $\underline{\beta}_{ij} = 0$ and $\underline{V}_{\beta_{ij}} = 10$, $i = 1, ..., n$, $j = 1, np + 2, ..., k_i$.

[25] The BVAR-GLP approach allows alternative prior variants, such as the sum-of-coefficients prior. We have estimated a number of these variants and, with the exception of the sum-of-coefficients prior, by and large the results do not change significantly. As expected with the stationary data we use, the sum-of-coefficients prior does not work particularly well.

estimates to predict the outcomes from 1985:Q2 to 1986:Q1. We proceed recursively in this fashion until 2015:Q4, thus generating a time series of point and density forecasts for each forecast horizon h, with $h = 1, ..., 4$.[26]

Next, for each of the seven key variables listed above we summarize the precision of the h-step-ahead point forecasts for model i, relative to that from a seven-variable VAR(p) benchmark, by means of the ratio of MSFEs:

$$MSFE_{ijh} = \frac{\sum_{\tau=\underline{t}}^{\bar{t}-h} e_{i,j,\tau+h}^{2}}{\sum_{\tau=\underline{t}}^{\bar{t}-h} e_{bcmk,j,\tau+h}^{2}}, \tag{32}$$

where the benchmark VAR(p) has flat prior and is estimated using OLS, $p = 5$, $\underline{t}$ and $\bar{t}$ denote the start and end of the out-of-sample period, and $e_{i,j,\tau+h}^{2}$ and $e_{bcmk,j,\tau+h}^{2}$ are the squared forecast errors of variable j at time τ and forecast horizon h associated with model i ($i \in$ {DFM,FAVAR,BVAR-BGR,BVAR-GLP,BVAR-IP,SSVS,N-J,SNS,N-G}) and the benchmark VAR(p) model, respectively. The point forecasts used to compute the forecast errors are obtained by averaging over the draws from the various models' h-step-ahead predictive densities. Values of $MSFE_{ijh}$ below one suggest that model i produces more accurate point forecasts than the VAR(p) benchmark for variable j and forecast horizon h.

We also assess the accuracy of the point forecasts of the various methods using the multivariate loss function of Christoffersen and Diebold (1998). Specifically, we compute the ratio between the multivariate weighted mean squared forecast error (WMSFE) of model i and the WMSFE of the benchmark VAR(p) model as follows:

$$WMSFE_{ih} = \frac{\sum_{\tau=\underline{t}}^{\bar{t}-h} we_{i,\tau+h}}{\sum_{\tau=\underline{t}}^{\bar{t}-h} we_{bcmk,\tau+h}}, \tag{33}$$

where $we_{i,\tau+h} = \left(e_{i,\tau+h}' \times W \times e_{i,\tau+h}\right)$ and $we_{bcmk,\tau+h} = \left(e_{bcmk,\tau+h}' \times W \times e_{bcmk,\tau+h}\right)$ are time $\tau + h$ weighted forecast errors of model i and the benchmark model, $e_{i,\tau+h}$ and $e_{bcmk,\tau+h}$ are the (7×1) vector of forecast errors for the key series we focus on, and W is a (7×7) matrix of weights. Following Carriero et al. (2011), we set the matrix W to be a diagonal matrix featuring on the diagonal the inverse of the variances of the series to be forecast.

As for the quality of the density forecasts, we follow Geweke and Amisano (2010) and compute the average log predictive likelihood differential between model i and the seven-variable VAR(p)

[26] Note that when $h > 1$, point forecasts are iterated and predictive simulation is used to produce the predictive densities.

benchmark,

$$ALPL_{ijh} = \frac{1}{\bar{t} - \underline{t} - h + 1} \sum_{\tau=\underline{t}}^{\bar{t}-h} \left(LPL_{i,j,\tau+h} - LPL_{bcmk,j,\tau+h} \right), \tag{34}$$

where $LPL_{i,j,\tau+h}$ ($LPL_{bcmk,j,\tau+h}$) denotes model i's (benchmark's) log predictive score of variable j, computed at time $\tau + h$, i.e., the log of the h-step-ahead predictive density evaluated at the outcome. Positive values of $ALPL_{ijh}$ indicate that for variable j and forecast horizon h on average model i produces more accurate density forecasts than the benchmark model.

Finally, in order to test the statistical significance of differences in point and density forecasts, we consider pairwise tests of equal predictive accuracy (henceforth, EPA; Diebold and Mariano, 1995; West, 1996) in terms of MSFE, WMSFE, and ALPL. All EPA tests we conduct are based on a two sided test with the null hypothesis being the seven-variable VAR(p) benchmark. We use standard normal critical values. Based on simulation evidence in Clark and McCracken (2013), when computing the variance estimator which enters the test statistic we rely on serial correlation robust standard errors, and incorporate the finite sample correction due to Harvey et al. (1997). In the tables, we use ***, ** and * to denote results which are significant at the 1%, 5% and 10% levels, respectively, in favor of the model listed at the top of each column.

6.3 Numerical accuracy of our proposed algorithm

Before we present the core of our results, we first compare the forecasts of our algorithm against those obtained using similar MCMC-based hierarchical models. In particular, we compare the performance of our Spike-and-Slab (SNS) approach to the MCMC-based variable selection algorithm of Kuo and Mallick (1998) and Korobilis (2013b). We also look at the relative performance of our Normal-Gamma (N-G) approach against the MCMC-based hierarchical Student-t prior algorithm described in Tipping (2001) and Polson and Scott (2010).[27] We denote the VAR with variable selection prior as MCMC-SNS, and the VAR with Student-t prior as MCMC-t.

[27] We estimate the BVAR with the Student-t shrinkage prior by using a mixture representation that places an independent Normal prior on the VAR coefficients and an Inverse-Gamma prior on each of their prior variances. As shown by Korobilis (2013a) in a univariate setting, the Normal-Inverse Gamma prior distribution is conditionally conjugate and leads to the use of a standard Gibbs sampler scheme. As Huber and Feldkircher (2017) show, a Normal-Gamma prior would instead require the use of a Metropolis-Hasting algorithm and much larger computational needs, rendering estimation prohibitively costly with large-dimensional VARs.

Table 1 shows, side to side, the relative forecast accuracy, as measured using the $WMSFE$ statistics, of the two MCMC-based hierarchical priors versus our N-G and SNS methods, across all VAR sizes we considered. In particular, the first two columns of this table compare the MCMC-t to our N-G prior, while the remaining two columns look at the comparison between MCMC-SNS and our SNS prior.[28] We begin by noting that due to the larger computational costs imposed by the MCMC-based algorithms, we could not successfully complete the estimation of the MCMC-SNS and MCMC-t methods in the X-large VAR case. In contrast, thanks to the simulation-free nature of our algorithm, we were able to carry out inference and forecasting using our two hierarchical prior variants for all VAR sizes. As for the relative accuracy of our estimation algorithm, the $WMSFE$ statistics in the Table show that the MCMC-t and our N-G hierarchical prior provide almost identical results, thus, confirming that our simulation-free approach is as accurate as its MCMC counterpart (at a fraction of the time). Regarding the two SNS approaches, we find that, at least for the Medium VAR case, our SNS prior produces results that are quite similar to those obtained using the MCMC-SNS algorithm. We also find that in the case of the large VAR, our approach appears to perform substantially better (more than 10% average improvement across all forecast horizons) than its MCMC counterpart. We attribute this result to the potential numerical instabilities that can plague the MCMC-based variable selection algorithms in high-dimensional settings. In fact, while both SNS algorithms require multiple evaluations of conditional likelihoods, similar to those described in equation (18), the MCMC-SNS algorithm will need to repeat such evaluations for each Monte Carlo iteration. In large dimensions, evaluation of the exponential VAR likelihood can result in overflow/underflow problems, and subsequent loss in numerical accuracy when computing the posterior inclusion probabilities. In this case our simulation-free SNS algorithm will likely be more stable than its MCMC analogue.

6.4 Forecasting results

Having established that our simulation-free hierarchical prior models are at least as precise as their MCMC equivalents at a fraction of the computing time, we now proceed to compare the performance of our methods to all the competing models we outlined in subsection 6.1. Table 2

[28] To perform this comparison, we have used identical prior moments throughout, or whenever this was not possible we followed the default choices and the recommendations in Korobilis (2013a,b).

provides a summary of the forecasting ability of each method by presenting its relative $WMSFE$ statistics. The table includes three panels, each one presenting results for a different VAR size, with the rows focusing on the various forecast horizons and the columns zooming in on the various methods we considered. We begin by noting that as it was the case with the analysis we presented in Table 1, in the case of the X-large VAR we are only able to report results for the seven models that do not rely on MCMC methods. In fact, despite our use of a High Performance Computing Cluster, we found that both the BVAR-IP and SSVS methods did not converge, either because of numerical instabilities or because the algorithm exceeded the total available resources.[29]

Next, looking across all three VAR dimensions and all four forecast horizons, we find that the hierarchical Spike-and-Slab (SNS) and Normal-Gamma (N-G) priors dominate all other methods in terms of forecasting accuracy, attaining the lowest $WMSFEs$ in 11 of the 12 cases considered in the table. As we saw in Figure 1, these two priors are very closely related so the numerical similarities of their $WMSFEs$ do not come as a surprise. Interestingly, while the forecasts from the improper Normal-Jeffreys (N-J) prior also tend to improve substantially over the benchmark seven-variable VAR, it appears that the N-J approach always lags behind our other two methods, especially as the forecast horizon increases. We attribute this result to the fact that the N-J is an improper prior, leading to an improper (and unbounded) posterior for the VAR parameters. Despite this, all our three hierarchical priors produce forecast gains that across the board are quite substantial, approaching or even exceeding 40% improvements in $WMSFE$ terms over the benchmark seven-variable VAR for a number of horizons. Gains relative to the alternative methods we considered are also in general quite large and significant, with a rough average improvement of 15-20% over the vast majority of the competing methods. The only exception to this rule is the BVAR-IP, which thanks to the use of the independent prior can fit the data better and produce forecasts that are only slightly inferior to the ones we obtain with our simulation-free method. However, it is worth pointing out that the BVAR-IP is the only BVAR method we considered that requires manual intervention in the tuning of the

[29] In principle, it could be possible to improve the computational efficiency of the MCMC algorithms we considered by splitting the MCMC chain into a number of parallel and shorter chains. However, there is really no speed-up available for the burn-in stage of the MCMC algorithm, as each chain must complete the full burn-in before generating draws that can be safely retained (see for example Geyer, 1992). While the severity of this issue depends on how quickly the sampler converges to its ergodic posterior distribution, generally speaking MCMC-based algorithms are incompatible with a fully-fledged parallelization.

overall shrinkage intensity parameter λ. While in this particular setting the recommendation of Sims and Zha (1998) of setting $\lambda = 0.2$ appears to work quite well, we have also found in our experimentation that many other (reasonable) values of λ yields considerably inferior forecasts for this method. The other drawback of the BVAR-IP method is that it relies on MCMC techniques, and as we discussed above does not adapt well to very high dimensional VARs.

Tables 3 to 5 present evidence on the performance of the various models for the seven variables of interest, relative to the benchmark seven-variable VAR. In particular, each table focuses on one specific VAR size, zooming into the relative $MSFE$ performance across the four forecast horizons and the seven variables of interest. Starting with Table 3, we find that in the case of the Medium VAR the BVAR-IP method is very competitive, especially at short horizons, while our SNS and N-G methods appear to hold a slight hedge over the alternative methods for the longer horizons. Table 4 show a similar pattern in the large VAR case, while in the X-large VAR, as Table 5 indicates, our methods generate the best $MSFEs$ in 16 of the 28 cases considered. The DFM and FAVAR follow right behind, with the DFM being particularly successful for GDP and the unemployment rate. Looking more specifically into the individual series, we find that the forecasts for CPI inflation never seem to outperform the benchmark VAR (as indicated by the Diebold-Mariano statistics), while in the case of the GDP deflator results for the various methods we considered are far better. Also, when moving to the to the X-large VAR, we find that GDP forecasts appear to dramatically improve for all horizons, and that the same holds to a smaller degree for employment. In summary, we find that our methods do very well, and even when they are not ranked first they tend to be very close to the best performing model(s).

Tables 6 to 8 repeat the same analysis by looking at the whole forecast distribution via the use of average log predictive likelihoods ($ALPL$s). Results appear more mixed in this case, with no single method emerging as a clear winner. Nevertheless, we find that our methods dominate in many instances, without ever falling too much behind in any individual case. We make three additional remarks. First, the BGR method seems to be performing extremely well for some series (mainly GDP, FEDFUNDS, GDPDEFL), when in the case of its point forecasts it wasn't among the top performing methods. Additionally, of interest is the quality of the density forecasts of CPI inflation for all methods other than the two relying on the natural conjugate prior. Improvements over the benchmark unrestricted VAR appear quite substantial, suggesting that even though none of the available methods improved in terms of point forecasts, when it comes to density forecasts the independent shrinkage priors are quite helpful. Finally, according to the Diebold-Mariano tests, a large proportion of $APLs$ for all variables and forecast horizons become statistically significant. This fact provides further assurances that larger information sets are useful in achieving sharper forecasts and controlling for forecast uncertainty.

7 Structural VARs and impulse response analysis

The excellent forecast performance of our methodology is in line with an expanding literature in statistics that praises the use of hierarchical priors for providing successful regularized estimation. As explained in Section 2, we have paired such priors with a fast approximate procedure that provides as output a joint parameter posterior $p\left(\boldsymbol{\beta}|\, \boldsymbol{y}\right)$ under the assumption that all the elements of the vector $\boldsymbol{\beta}$ are a-posteriori uncorrelated. This approximation appears to be quite satisfactory in the high-dimensional forecasting application we have considered, where the final outcome of interest is simply a set of predictions for some economic variables of interest.

In addition to forecasting, VARs are also used regularly to identify structural shocks and assess the transmission mechanisms of the macro-economy through impulse response analysis and historical decompositions. In these cases, the assumption of a-posteriori independence may hinder the ability of the economist to provide reliable policy recommendations. In this section, we present a simple modification of our algorithm that is better suited for structural analysis.

In order to demonstrate this procedure, we follow papers such as Giannone et al. (2015) and generate 500 artificial datasets of $T = 216$ quarters from a large-scale dynamic stochastic general equilibrium (DSGE) model. The model we use is an extension of Görtz and Tsoukalas (2017) and Görtz et al. (2016), and focuses on sectoral total factor productivity (TFP) shocks and financial frictions.[30] Among all possible sectoral and aggregate variables that this model generates, we focus only on the aggregate ones, to stay consistent with the bulk of the news shock literature.[31] In particular, we follow Barsky and Sims (2011) and use TFP, real GDP, consumption, and hours as our four variables of interest; in addition, to better identify news shocks, we include three

[30] More specifically, we generate the artificial data using the default parameter settings that Görtz et al. (2016) use when financial frictions are present.

[31] See Beaudry and Portier (2013) for an excellent review of empirical studies on news and business cycles.

additional series from the DSGE model, namely inflation, interest rate spread (the difference between long-term and short-term interest rates), and equity prices.[32] Finally, as the news shocks are not directly observed in a VAR setting, we rely on the identification scheme of Forni et al. (2014) to extract them.[33]

For each of the 500 datasets, we use the artificial data on the seven variables listed above to estimate a VAR with flat priors and a hierarchical prior BVAR. In particular, to estimate the latter model we rely on a simple two-stage procedure. In the first step of this procedure, we use the estimation algorithm described in Section 4 along with the hierarchical Spike-and-Slab prior to obtain posterior inclusion probabilities $\widehat{\pi}_{ij}$ for each of the VAR coefficients. Next, in the second step, we insert the restrictions implied by the posterior inclusion probabilities in a BVAR, which is estimated using an independent Normal-Wishart prior.[34]

Figure 6 plots the DSGE theoretical impulse responses to a productivity news shock, along with the average across the 500 replications of the median impulse responses for the flat prior (VAR) and our hierarchical prior (BVAR).[35] In general, both the VAR and BVAR models seem to capture fairly well the responses of output, consumption and hours. On the other hand, news shock in the DSGE model are anticipated 12 quarters ahead, therefore the response of TFP is zero for the first 12 periods. Such feature is generally harder to capture with a VAR or BVAR. Nevertheless, the empirical responses of TFP of both models are still quite reasonable, and in line with the VAR estimates reported elsewhere (Barsky and Sims, 2011). Next, Figure 7 provides a more accurate assessment of the differences in the estimated impulse responses. For

[32] Forni and Gambetti (2014) have shown that many of the smaller VARs considered in this literature are non-fundamental, meaning that they will not recover news shocks correctly. On the other hand, Beaudry et al. (2015) have argued that even non-fundamental VARs can correctly recover the responses of TFP to news shock. Regardless of this, larger information sets are still needed in order to identify correctly the remaining responses of interest to policy-makers, namely, output, consumption and hours.

[33] The identification scheme of Forni et al. (2014) relies on a combination of long and short-run restrictions on TFP. The alternative identification schemes proposed in Barsky and Sims (2011) and Francis et al. (2014) produce identical results.

[34] In particular, we start from (24) and rewrite the VAR in (21) in its SUR form. Using notation from Koop and Korobilis (2010), we rewrite the reduced-form VAR in (21) as $\boldsymbol{Y} = \widetilde{\boldsymbol{X}}\boldsymbol{B} + \boldsymbol{V}$ where $\boldsymbol{Y} = (\boldsymbol{y}_1', ..., \boldsymbol{y}_n')'$ and $\boldsymbol{V} = (\boldsymbol{v}_1', ..., \boldsymbol{v}_n')'$ are $Tn \times 1$ vectors, while $\widetilde{\boldsymbol{X}}$ is a $Tn \times K$ block-diagonal matrix obtained by stacking together the $T \times k_i$ matrices $\widetilde{\boldsymbol{X}}_1$,...,$\widetilde{\boldsymbol{X}}_n$ that incorporate the constraints implied by the estimated PIPs in (19). The elements in the generic matrix $\widetilde{\boldsymbol{X}}_i$ $(i = 1, ..., n)$, in turn, are computed by multiplying each row of $\boldsymbol{X}_i$ by $\widehat{\boldsymbol{\pi}}_i$, the $k_i \times 1$ vector of PIPs estimated from the VAR's i-th equation, i.e. $\widetilde{\boldsymbol{X}}_{i,t} = \boldsymbol{X}_{i,t} \circ \widehat{\boldsymbol{\pi}}_i'$, where $\circ$ denotes the Hadamard product, and $t = 1, ..., T$.

[35] Interestingly, the shape of the responses of output and consumption have a distinct double-hump shape. This is the direct consequence of working with a model with financial frictions; see Figure 10 of Görtz et al. (2016) for more details.

each of the 500 replications, we compute the difference between the theoretical DSGE response and the estimated VAR and BVAR median responses, across the seven variables and the 40 horizons. Then, for each variable and horizon, we take the average of the squared errors across replications (MSE). Figure 7 plots the ratio between the MSE of the VAR with flat priors and the MSE of the hierarchical BVAR. As it can be seen from the figure, for the vast majority of periods the MSE ratios are higher than one, implying that the two-step BVAR procedure based on the hierarchical Spike-and-Slab prior generates more accurate responses than the flat prior VAR.

8 Conclusions

We have introduced a novel methodology for estimating BVARs, which features a number of desirable properties including flexible priors, closed-form posterior moments, and large computational efficiency. We exploited the flexibility of this novel approach to study empirically the benefits of a wide class of hierarchical shrinkage priors that lead to individualized adaptive shrinkage on the VAR coefficients. Thanks to the estimation method we introduced, we are able to derive analytical expressions for the marginal posteriors implied by three popular cases of hierarchical priors, namely Normal-Jeffreys, Spike-and-Slab, and Normal-Gamma. Our approach works extremely well with BVARs of both medium and large dimensions, delivering analytical approximations to the marginal posterior distributions of the BVAR coefficients that are very accurate. In addition, our proposed algorithm for posterior inference is multiple times faster than existing Bayesian VAR methods that rely on simulation methods. We implement a thorough Monte Carlo analysis to quantify the benefits of our approach, and find that it can recover very accurately the underlying VAR coefficients. We also demonstrate, using an extensive forecasting application with VARs of up to 124 equations, the benefits of our adaptive shrinkage procedure in preventing over-fitting of large VARs and providing excellent forecasting performance. Finally, we demonstrate using a simulated numerical example with artificial data extracted from a large structural macroeconomic model, that our algorithm can be useful also in recovering structural impulse responses.

References

Banbura, M., D. Giannone, and L. Reichlin (2010): "Large Bayesian Vector Autoregressions," *Journal of Applied Econometrics*, 25, 71–92.

Barber, D. (2012): *Bayesian Reasoning and Machine Learning*, Cambridge University Press.

Barsky, R. B. and E. R. Sims (2011): "News shocks and business cycles," *Journal of Monetary Economics*, 58, 273 – 289.

Beaudry, P., P. Fve, A. Guay, and F. Portier (2015): "When is Nonfundamentalness in VARs a Real Problem? An Application to News Shocks," Working Paper 21466, National Bureau of Economic Research.

Beaudry, P. and F. Portier (2013): "News Driven Business Cycles: Insights and Challenges," Working Paper 19411, National Bureau of Economic Research.

Bernanke, B. S., J. Boivin, and P. Eliasz (2005): "Measuring the effects of monetary policy: A factor-augmented vector autoregressive (favar) approach," *The Quarterly Journal of Economics*, 120, 387–422.

Carriero, A., T. Clark, and M. Marcellino (2017): "Large Vector Autoregressions with stochastic volatility and flexible priors," Working paper 16-17, Federal Reserve Bank of Cleveland.

Carriero, A., G. Kapetanios, and M. Marcellino (2011): "Forecasting large datasets with Bayesian reduced rank multivariate models," *Journal of Applied Econometrics*, 26, 735–761.

——— (2012): "Forecasting government bond yields with large Bayesian vector autoregressions," *Journal of Banking and Finance*, 36, 2026 – 2047.

Christoffersen, P. F. and F. X. Diebold (1998): "Cointegration and Long-Horizon Forecasting," *Journal of Business & Economic Statistics*, 16, 450–458.

Clark, T. and M. McCracken (2013): "Advances in Forecast Evaluation," in *Handbook of Economic Forecasting*, ed. by A. Timmermann and G. Elliott, Elsevier, vol. 2, chap. 20, 1107–1201.

Del Negro, M. and F. Schorfheide (2004): "Priors from General Equilibrium Models for VARS," *International Economic Review*, 45, 643–673.

Diebold, F. X. and R. S. Mariano (1995): "Comparing Predictive Accuracy," *Journal of Business & Economic Statistics*, 13, 253–263.

Doan, T., R. Litterman, and C. Sims (1984): "Forecasting and conditional projection using realistic prior distributions," *Econometric Reviews*, 3, 1–100.

Figueiredo, M. A. T. (2003): "Adaptive sparseness for supervised learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25, 1150–1159.

Forni, M. and L. Gambetti (2014): "Sufficient information in structural VARs," *Journal of Monetary Economics*, 66, 124–136.

Forni, M., L. Gambetti, and L. Sala (2014): "No News in Business Cycles," *The Economic Journal*, 124, 1168–1191.

Francis, N., M. T. Owyang, J. E. Roush, and R. DiCecio (2014): "A Flexible Finite-Horizon Alternative to Long-Run Restrictions with an Application to Technology Shocks," *The Review of Economics and Statistics*, 96, 638–647.

Frisch, R. and F. V. Waugh (1933): "Partial Time Regressions as Compared with Individual Trends," *Econometrica*, 1, 387–401.

George, E. I., D. Sun, and S. Ni (2008): "Bayesian stochastic search for VAR model restrictions," *Journal of Econometrics*, 142, 553–580.

Geweke, J. and G. Amisano (2010): "Comparing and evaluating Bayesian predictive distributions of asset returns," *International Journal of Forecasting*, 26, 216 – 230.

Geyer, C. J. (1992): "Practical Markov Chain Monte Carlo," *Statist. Sci.*, 7, 473–483.

Giannone, D., M. Lenza, and G. E. Primiceri (2015): "Prior Selection for Vector Autoregressions," *Review of Economics and Statistics*, 97, 436–451.

Görtz, C. and J. D. Tsoukalas (2017): "News and Financial Intermediation in Aggregate Fluctuations," *The Review of Economics and Statistics*, 99, 514–530.

Görtz, C., J. D. Tsoukalas, and F. Zanetti (2016): "News Shocks under Financial Frictions," .

Griffin, J. E. and P. J. Brown (2010): "Inference with normal-gamma prior distributions in regression problems," *Bayesian Anal.*, 5, 171–188.

Harvey, D., S. Leybourne, and P. Newbold (1997): "Testing the equality of prediction mean squared errors," *International Journal of Forecasting*, 13, 281 – 291.

Hausman, J. A. (1983): "Chapter 7 Specification and estimation of simultaneous equation models," *Handbook of Econometrics*, 1, 391 – 448.

Hobert, J. P. and G. Casella (1996): "The Effect of Improper Priors on Gibbs Sampling in Hierarchical Linear Mixed Models," *Journal of the American Statistical Association*, 91, 1461–1473.

Huber, F. and M. Feldkircher (2017): "Adaptive Shrinkage in Bayesian Vector Autoregressive Models," *Journal of Business and Economic Statistics*, forthcoming.

Kadiyala, K. R. and S. Karlsson (1997): "Numerical methods for estimation and inference in Bayesian VAR-models," *Journal of Applied Econometrics*, 12, 99–132.

Koop, G. (2003): *Bayesian Econometrics*, John Wiley & Sons, Ltd.

Koop, G. and D. Korobilis (2010): "Bayesian Multivariate Time Series Methods for Empirical Macroeconomics," *Foundations and Trends in Econometrics*, 3, 267–358.

Koop, G., D. Korobilis, and D. Pettenuzzo (2017): "Bayesian Compressed Vector Autoregressions," *Journal of Econometrics*, forthcoming, working paper.

Korobilis, D. (2013a): "Hierarchical shrinkage priors for dynamic regressions with many predictors," *International Journal of Forecasting*, 29, 43–59.

——— (2013b): "VAR Forecasting Using Bayesian Variable Selection," *Journal of Applied Econometrics*, 28, 204–230.

Kuo, L. and B. Mallick (1998): "Variable Selection for Regression Models," *The Indian Journal of Statistics, Series B (1960-2002)*, 60, 65–81.

Litterman, R. B. (1979): "Techniques of forecasting using vector autoregressions," Working Papers 115, Federal Reserve Bank of Minneapolis.

Mitchell, T. J. and J. J. Beauchamp (1988): "Bayesian Variable Selection in Linear Regression," *Journal of the American Statistical Association*, 83, 1023–1032.

Polson, N. and J. Scott (2010): "Shrink Globally, Act Locally: Sparse Bayesian Regularization and Prediction," in *Bayesian Statistics*, Oxford University Press, vol. 9, 1–24.

Primiceri, G. E. (2005): "Time Varying Structural Vector Autoregressions and Monetary Policy," *The Review of Economic Studies*, 72, 821–852.

Sims, C. A. and T. Zha (1998): "Bayesian Methods for Dynamic Multivariate Models," *International Economic Review*, 39, 949–968.

Smith, M. and R. Kohn (2002): "Parsimonious Covariance Matrix Estimation for Longitudinal Data," *Journal of the American Statistical Association*, 97, 1141–1153.

Stock, J. H. and M. W. Watson (2002): "Macroeconomic Forecasting Using Diffusion Indexes," *Journal of Business & Economic Statistics*, 20, 147–162.

Tibshirani, R. (1996): "Regression Shrinkage and Selection Via the Lasso," *Journal of the Royal Statistical Society, Series B*, 58, 267–288.

Tipping, M. E. (2001): "Sparse Bayesian Learning and the Relevance Vector Machine," *Journal of Machine Learing Research*, 1, 211–244.

van den Boom, W., G. Reeves, and D. B. Dunson (2015a): "Quantifying Uncertainty in Variable Selection with Arbitrary Matrices," in *Proceedings of the IEEE International*

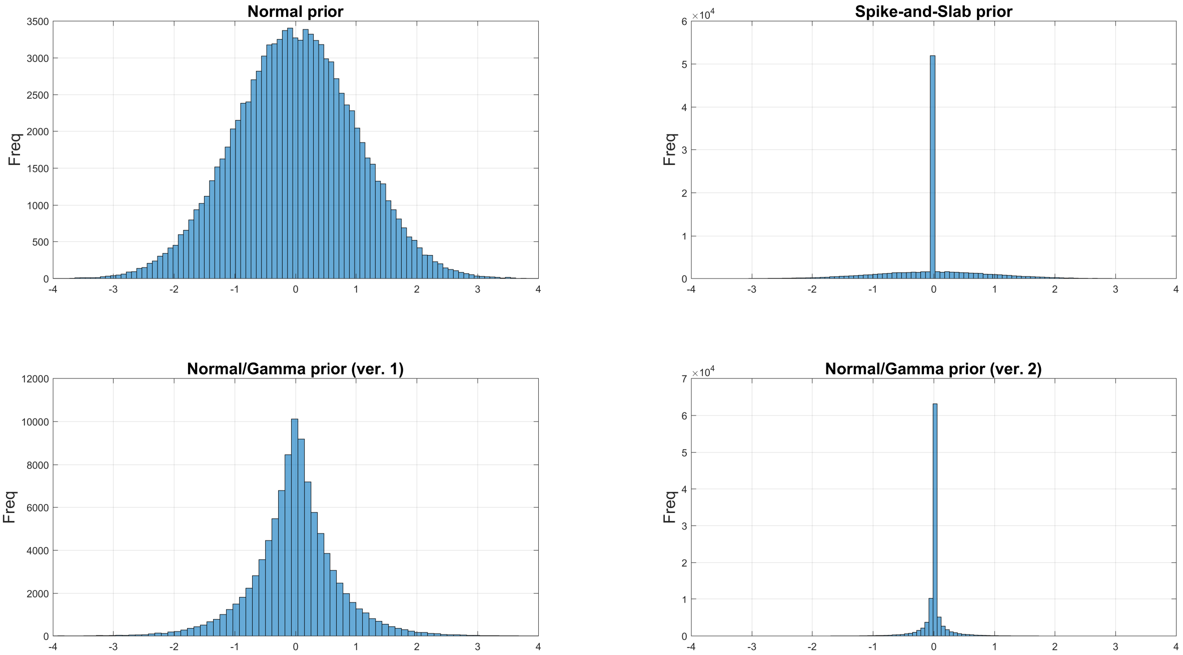
Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP), Cancun, Mexico.

——— (2015b): "Scalable Approximations of Marginal Posteriors in Variable Selection," Working paper, Duke University Department of Statistical Science. Available at arXiv:1506.06629.

West, K. D. (1996): "Asymptotic Inference about Predictive Ability," *Econometrica*, 64, pp. 1067–1084.

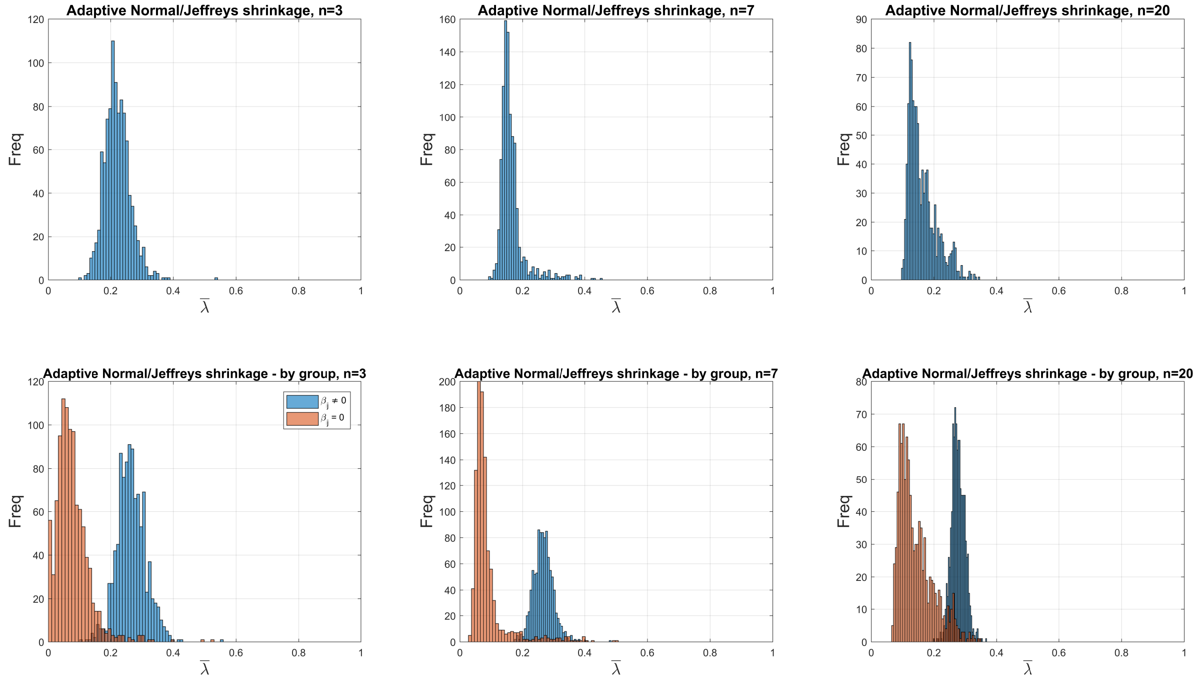
Figures and Tables

Figure 1. Histograms of hierarhical priors



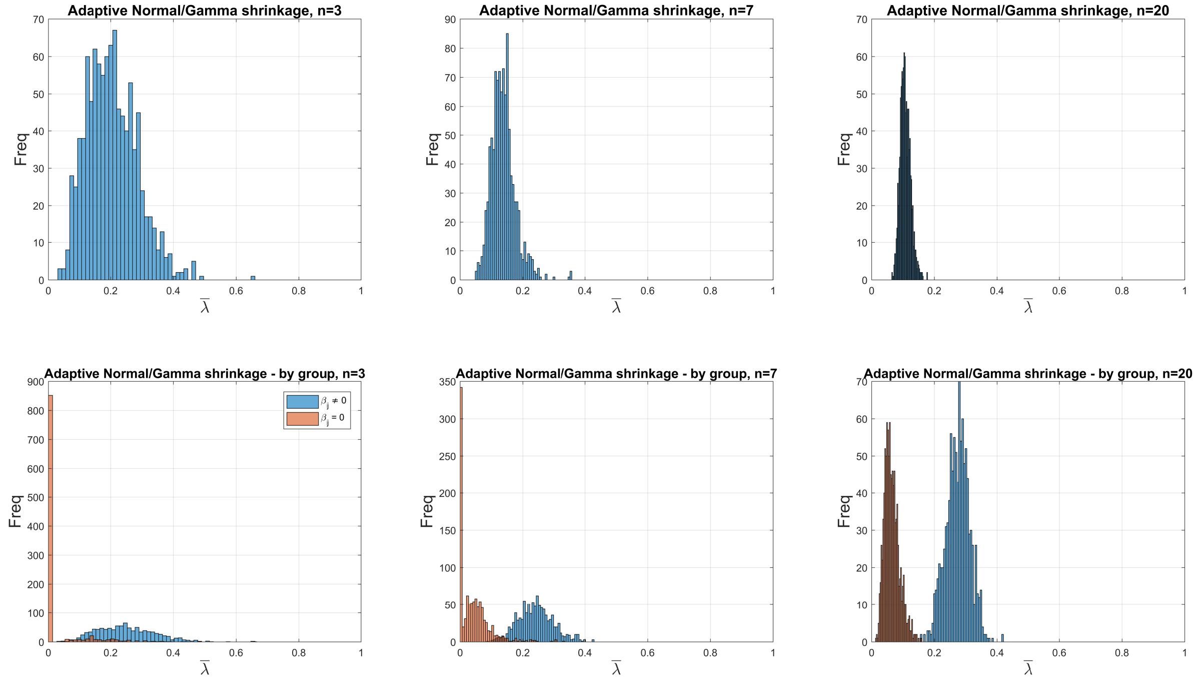
Top left panel: an example of a Normal prior for β_j in one dimension, where $\beta_j \sim \mathcal{N}\left(0, \underline{V}_{\beta_j}\right)$, and $\underline{V}_{\beta_j} = 10$. Top right panel: an example of a Spike-and-Slab prior for β_j in one dimension, where $\beta_j \sim (1-\lambda_j)\,\delta_0 + \lambda_j \mathcal{N}\left(0, \underline{V}_{\beta_j}\right)$, $\lambda_j \sim Bernoulli\left(\underline{\pi}_0\right)$, and $\underline{\pi}_0 = 0.5$. Bottom panels: two examples of a hierarchical Normal/Gamma prior for β_j in one dimension, where the hyperparameter λ_j^2 has been integrated out, i.e. $p\left(\beta_j\right) = \int p\left(\beta_j|\lambda_j^2\right) p\left(\lambda_j^2\right) d\lambda_j^2$, with $\beta_j|\lambda_j^2 \sim \mathcal{N}\left(0, \lambda_j^2 \underline{V}_{\beta_j}\right)$ and $\lambda_j^2 \sim \mathcal{G}\left(\underline{c}_1, \underline{c}_2\right)$. In the bottom left panel, we set $\underline{c}_1 = 1$ $\underline{c}_2 = 2$, while in the bottom right panel we have $\underline{c}_1 = 0.1$ $\underline{c}_2 = 2$.

Figure 2. Monte Carlo simulation - Shrinkage intensity, Normal/Jeffreys prior



The top three panels plot the empirical distribution of the average estimated shrinkage intensity $\overline{\lambda} = \frac{1}{K}\sum_{i=1}^{n}\sum_{j=1}^{k_i}\widehat{\lambda}_{ij}$ for $n = 3$, $n = 7$, and $n = 20$-variable VAR(p), averaged over all VAR coefficients. $K = \sum_{i=1}^{n} k_i$ denotes the total number of VAR coefficients, including the covariance terms in $\mathbf{\Phi}$, and $k_i = np + i$. Results are based on our adaptive shrinkage procedure and the Normal/Jeffreys prior. The bottom three panels plot the average shrinkage intensity estimated by our adaptive procedure, broken down according to whether the corresponding VAR coefficients in the simulated data are equal to zero (red bars) or not (blue bars). All empirical distributions are obtained by simulating $1,000$ VAR(p) of sample size $T = 150$ and lag length $p = 2$. See Section 5 for additional details on the design of the Monte Carlo simulation.

Figure 3. Monte Carlo simulation - Shrinkage intensity, Normal/Gamma prior



Adaptive Normal/Gamma shrinkage, n=3

Adaptive Normal/Gamma shrinkage, n=7

Adaptive Normal/Gamma shrinkage, n=20

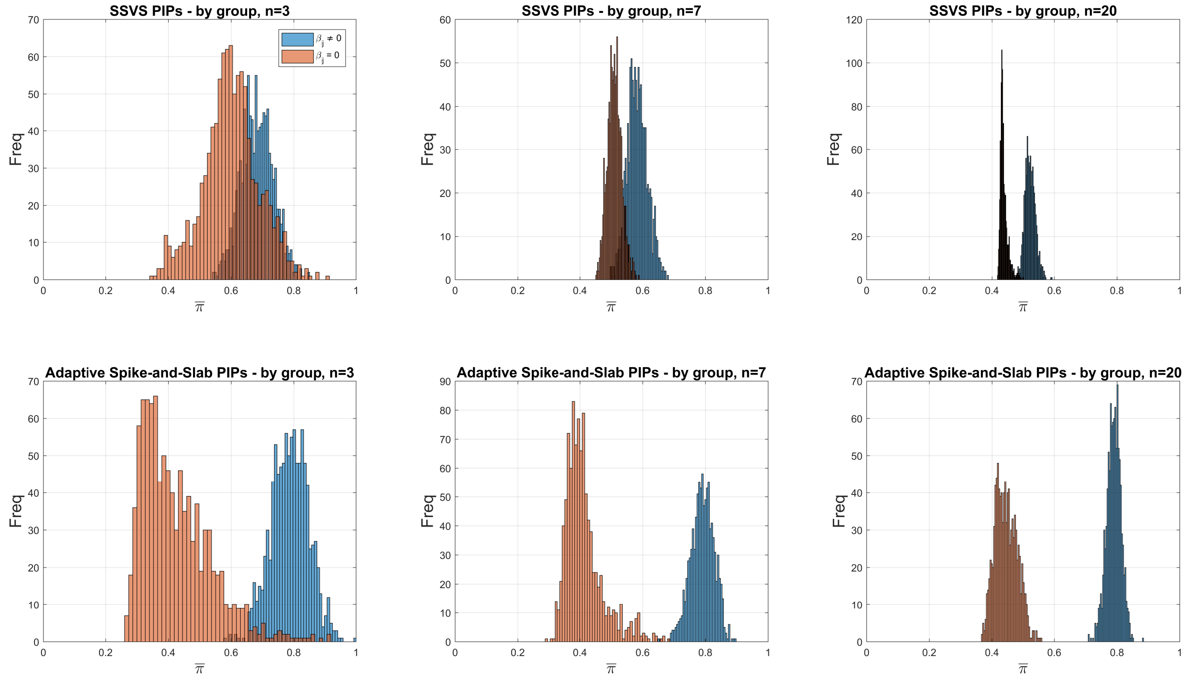
Adaptive Normal/Gamma shrinkage - by group, n=3

Adaptive Normal/Gamma shrinkage - by group, n=7

Adaptive Normal/Gamma shrinkage - by group, n=20

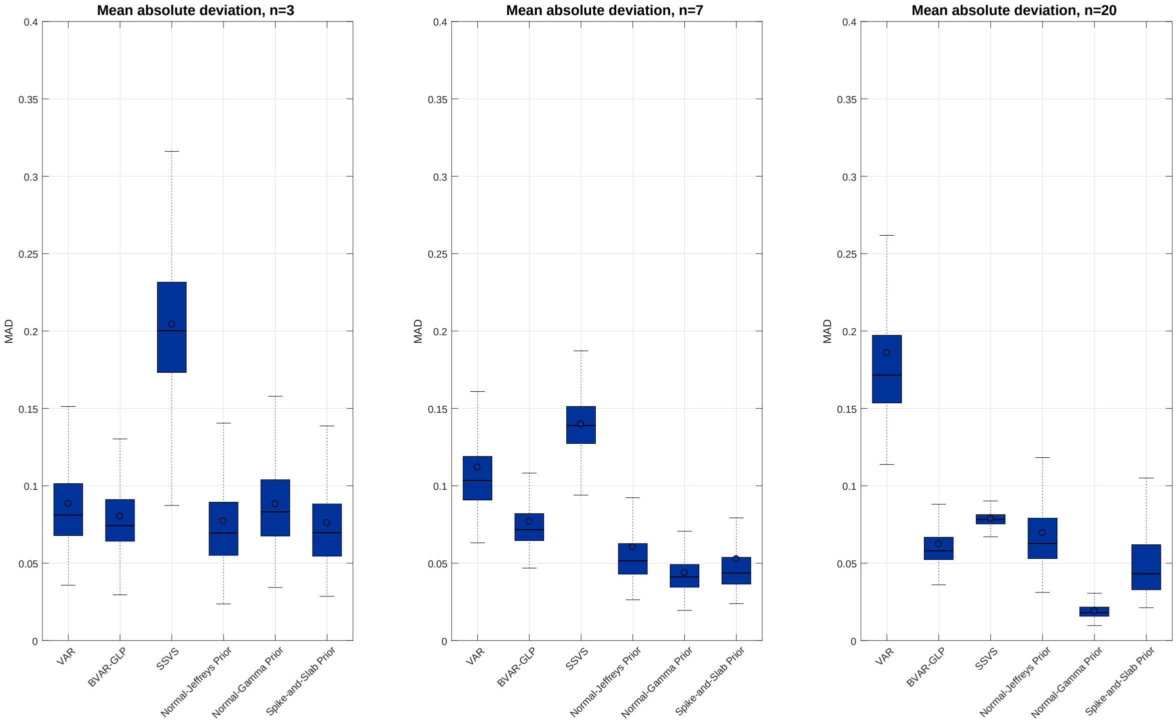
The top three panels plot the empirical distribution of the average estimated shrinkage intensity $\overline{\lambda} = \frac{1}{K}\sum_{i=1}^{n}\sum_{j=1}^{k_i}\widehat{\lambda}_{ij}$ for $n = 3$, $n = 7$, and $n = 20$-variable VAR(p), averaged over all VAR coefficients. $K = \sum_{i=1}^{n} k_i$ denotes the total number of VAR coefficients, including the covariance terms in $\mathbf{\Phi}$, and $k_i = np + i$. Results are based on our adaptive shrinkage procedure and the Normal/Gamma prior. The bottom three panels plot the average shrinkage intensity estimated by our adaptive procedure, broken down according to whether the corresponding VAR coefficients in the simulated data are equal to zero (red bars) or not (blue bars). All empirical distributions are obtained by simulating $1,000$ VAR(p) of sample size $T = 150$ and lag length $p = 2$. See Section 5 for additional details on the design of the Monte Carlo simulation.

Figure 4. Monte Carlo simulation - Posterior Inclusion Probabilities (PIPs)



The top three panels of this figure plot the empirical distribution of the average posterior inclusion probability (PIP) obtained using the George et al. (2008) SSVS approach for $n = 3$, $n = 7$, and $n = 20$-variable VAR(p), and broken down according to whether the corresponding VAR coefficients in the simulated data are equal to zero (red bars) or not (blue bars). The bottom three panels plot the analogous empirical distributions of the averaged PIPs estimated using our adaptive shrinkage procedure with the Spike-and-Slab prior. All empirical distributions are obtained by simulating $1,000$ VAR(p) of sample size $T = 150$ and lag length $p = 2$. See Section 5 for additional details on the design of the Monte Carlo simulation.

Figure 5. Monte Carlo simulation - Mean Absolute Deviations

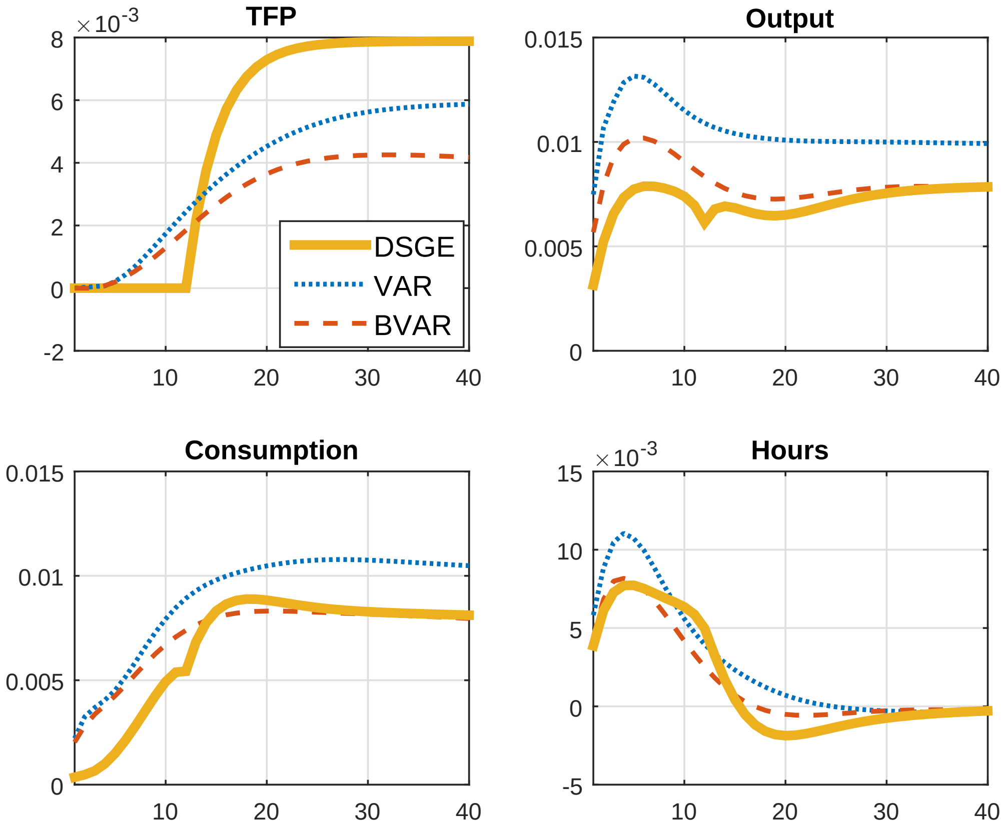


This figure reports box plots for the empirical distributions of the Mean Absolute Deviations (MAD), obtained from estimating a VAR(p) with OLS, a BVAR using the Giannone et al. (2015) (BVAR-GLP), the George et al. (2008) SSVS approach, and our adaptive shrinkage procedure with Normal/Jeffreys, Normal/Gamma, and Spike-and-Slab priors. These empirical distributions are obtained by simulating $1,000$ VAR(p) of sample size $T = 150$ and lag length $p = 2$. For each of the approaches listed and each of the 1,000 simulations we compute the Mean Absolute Deviation (MAD), defined as

$$MAD^{(r,s)} = \frac{1}{K}\sum_{i=1}^{n}\sum_{j=1}^{k_i}\left|\beta_{ij}^{(r)} - \widehat{\beta}_{ij}^{(r,s)}\right|,$$

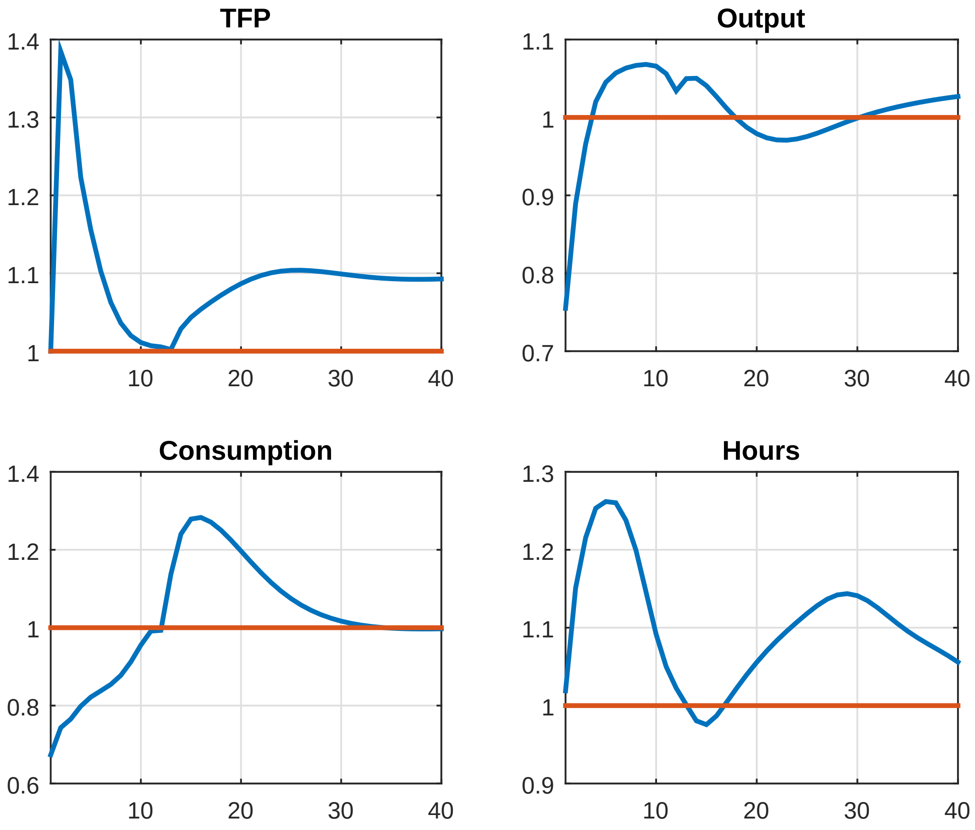
where s denotes the method used, $r = 1, ..., 1,000$ keeps track of the MC simulations, $K = \sum_{i=1}^{n} k_i$ denotes the total number of lag coefficients in the VAR, $\beta_{ij}^{(r)}$ is the true DGP coefficient from the r-th simulation, and $\widehat{\beta}_{ij}^{(r,s)}$ denotes the (posterior mean of the) corresponding estimate according to method s. Results are reported separately for $n = 3$, $n = 7$, and $n = 20$-variable VARs.

Figure 6. Impulse responses on simulated data



This figure reports the impulse responses to a productivity news shock in the DSGE model used to generate the data (solid line), and the median across Monte Carlo replications of the BVAR (dashed line) and the VAR (dotted line) impulse responses.

Figure 7. Ratio of MSE: VAR versus BVAR



This figure reports the ratio of the MSE of the VAR over the MSE of the BVAR. Values larger than one indicate that the MSE of the VAR is larger than that of the BVAR.

Table 1. Out-of-sample point forecast performance of our hierarchical BVARs against equivalent specifications estimated using MCMC: Multivariate results

		Medium VAR		
Forc. h	*MCMC-t*	*N-G*	*MCMC-SNS*	*SNS*
h=1	0.587***	0.587***	0.619***	0.607***
h=2	0.650***	0.647***	0.664***	0.657***
h=3	0.725***	0.707***	0.734***	0.720***
h=4	0.741***	0.715***	0.744***	0.736***
		Large VAR		
	MCMC-t	*N-G*	*MCMC-SNS*	*SNS*
h=1	0.579***	0.583***	0.838	0.606***
h=2	0.657***	0.646***	0.749**	0.635***
h=3	0.715***	0.704***	0.789**	0.694***
h=4	0.738***	0.719***	0.821**	0.710***
		X-large VAR		
	MCMC-t	*N-G*	*MCMC-SNS*	*SNS*
h=1	–	0.591***	–	0.621***
h=2	–	0.646***	–	0.651***
h=3	–	0.703***	–	0.705***
h=4	–	0.723***	–	0.722***

This table reports the ratio between the multivariate weighted mean squared forecast error (WMSFE) of model i and the WMSFE of the benchmark seven-variable VAR(p) model, computed as

$$WMSFE_{ih} = \frac{\sum_{\tau=\underline{t}}^{\overline{t}-h} we_{i,\tau+h}}{\sum_{\tau=\underline{t}}^{\overline{t}-h} we_{bcmk,\tau+h}},$$

where $we_{i,\tau+h} = \left(e'_{i,\tau+h} \times W \times e_{i,\tau+h}\right)$ and $we_{bcmk,\tau+h} = \left(e'_{bcmk,\tau+h} \times W \times e_{bcmk,\tau+h}\right)$ denote the weighted forecast errors of model i and the benchmark model at time $\tau + h$, $e_{i,\tau+h}$ and $e_{bcmk,\tau+h}$ are the $(N \times 1)$ vector of forecast errors, and W is an $(N \times N)$ matrix of weights. Throughout the table, we focus on $N = 7$ and the following series {PAYEMS, CPIAUCSL,FEDFUNDS, GDP, UNRATE, GDPDEFL, GS10} and set W to be a diagonal matrix featuring on the diagonal the inverse of the variances of the series to be forecast. $\underline{t}$ and $\overline{t}$ denote the start and end of the out-of-sample period, $i \in$ {MCMC-t, MCMC-SNS, SNS, N-G}, and $h = 1, ..., 4$. All forecasts are generated out-of-sample using recursive estimates of the models, with the out of sample period starting in 1985:Q1 and ending in 2015:Q4. * significance at the 10% level; ** significance at the 5% level; *** significance at the 1% level.

Table 2. Out-of-sample point forecast performance: Multivariate results

					Medium VAR				
Forc. h	*DFM*	*FAVAR*	*BVAR-BGR*	*BVAR-GLP*	*BVAR-IP*	*SSVS*	*N-J*	*SNS*	*N-G*
h=1	0.816*	0.723***	0.816***	0.775***	0.592***	0.876	0.624***	0.607***	**0.587*** **
h=2	0.803**	0.699***	0.901	0.913	**0.643*** **	0.754**	0.790***	0.657***	0.647***
h=3	0.811**	0.743***	0.872*	0.890	0.719***	0.799**	0.884	0.720***	**0.707*** **
h=4	0.781**	0.748***	0.838**	0.851*	0.736***	0.772**	0.895	0.736***	**0.715*** **
					Large VAR				
	DFM	*FAVAR*	*BVAR-BGR*	*BVAR-GLP*	*BVAR-IP*	*SSVS*	*N-J*	*SNS*	*N-G*
h=1	0.781**	0.686***	0.776***	0.797**	0.589***	0.691**	0.608***	0.606***	**0.583*** **
h=2	0.805**	0.711***	0.864*	0.900*	0.653***	0.731**	0.694***	**0.635*** **	0.646***
h=3	0.808***	0.739***	0.828**	0.860**	0.726***	0.768***	0.761***	**0.694*** **	0.704***
h=4	0.796**	0.759***	0.836**	0.849**	0.749***	0.768**	0.775**	**0.710*** **	0.719***
					X-large VAR				
	DFM	*FAVAR*	*BVAR-BGR*	*BVAR-GLP*	*BVAR-IP*	*SSVS*	*N-J*	*SNS*	*N-G*
h=1	0.790*	0.809**	0.716***	0.918	–	–	0.615***	0.621***	**0.591*** **
h=2	0.767***	0.782***	0.803***	0.846*	–	–	0.698***	0.651***	**0.646*** **
h=3	0.726***	0.742***	0.764***	0.836**	–	–	0.761***	0.705***	**0.703*** **
h=4	0.739***	0.771***	0.769***	0.866**	–	–	0.798**	**0.722*** **	0.723***

This table reports the ratio between the multivariate weighted mean squared forecast error (WMSFE) of model i and the WMSFE of the benchmark seven-variable VAR(p) model, computed as

$$WMSFE_{ih} = \frac{\sum_{\tau=\underline{t}}^{\bar{t}-h} we_{i,\tau+h}}{\sum_{\tau=\underline{t}}^{\bar{t}-h} we_{bcmk,\tau+h}},$$

where $we_{i,\tau+h} = \left(e'_{i,\tau+h} \times W \times e_{i,\tau+h}\right)$ and $we_{bcmk,\tau+h} = \left(e'_{bcmk,\tau+h} \times W \times e_{bcmk,\tau+h}\right)$ denote the weighted forecast errors of model i and the benchmark model at time $\tau + h$, $e_{i,\tau+h}$ and $e_{bcmk,\tau+h}$ are the $(N \times 1)$ vector of forecast errors, and W is an $(N \times N)$ matrix of weights. Throughout the table, we focus on $N = 7$ and the following series {PAYEMS, CPIAUCSL,FEDFUNDS, GDP, UNRATE, GDPDEFL, GS10} and set W to be a diagonal matrix featuring on the diagonal the inverse of the variances of the series to be forecast. $\underline{t}$ and $\bar{t}$ denote the start and end of the out-of-sample period, $i \in$ {DFM, FAVAR, BVAR-BGR, BVAR-GLP, BVAR-IP, SSVS, N-J, SNS, N-G}, and $h = 1, ..., 4$. All forecasts are generated out-of-sample using recursive estimates of the models, with the out of sample period starting in 1985:Q1 and ending in 2015:Q4. Bold numbers indicate the lowest WMSFE across all models for any given VAR size - forecast horizon pair. * significance at the 10% level; ** significance at the 5% level; *** significance at the 1% level.

Table 3. Out-of-sample point forecast performance, Medium VAR

Variable	*DFM*	*FAVAR*	*BVAR-BGR*	*BVAR-GLP*	*BVAR-IP*	*SSVS*	*N-J*	*SNS*	*N-G*
					$h = 1$				
PAYEMS	1.292	0.627**	0.706*	0.687**	0.548**	0.929	0.561*	__0.485**__	0.489**
CPIAUCSL	1.262	1.033	0.959	__0.902__	0.992	1.104	1.000	0.944	0.931
FEDFUNDS	0.452**	0.488***	0.704**	0.700**	__0.275**__	0.329**	0.339***	0.321**	0.283**
GDP	1.085	0.818*	0.859	0.788*	0.719**	1.571	__0.706**__	0.708**	0.741**
UNRATE	0.815	0.824	0.809	0.884	__0.642*__	1.622	0.763	0.655	0.646*
GDPDEFL	0.920	0.862	0.901	0.833*	0.821	1.004	__0.763**__	0.786*	0.770**
GS10	0.800*	0.775**	0.880	0.753***	__0.664***__	0.701***	0.711***	0.739**	0.687***
					$h = 2$				
PAYEMS	0.882	0.608**	0.855	0.851	0.536**	0.835	0.707	0.540**	__0.478***__
CPIAUCSL	1.031	0.997	0.983	1.011	__0.934__	0.957	1.056	0.971	0.967
FEDFUNDS	0.499***	__0.356***__	0.773**	0.847	0.380***	0.390***	0.554***	0.381***	0.375***
GDP	1.095	0.884	0.962	0.987	__0.704**__	1.024	0.853	0.718**	0.724**
UNRATE	0.794	0.743*	0.944	0.978	__0.695**__	0.913	0.943	0.754	0.699**
GDPDEFL	__0.790**__	0.924	0.934	0.824**	0.797**	0.815*	0.879	0.812*	0.824*
GS10	0.872	0.787**	0.997	0.961	0.786*	0.786**	0.857*	__0.777**__	0.791**
					$h = 3$				
PAYEMS	0.775	0.635**	0.847	0.862	0.617**	0.793	0.851	0.627**	__0.534***__
CPIAUCSL	0.962	0.945	0.994	1.017	0.945	__0.941__	1.042	0.974	0.962
FEDFUNDS	0.590***	0.517***	0.731**	0.785**	0.528***	0.527***	0.719	__0.507***__	0.526***
GDP	0.991	0.861	0.901	0.939	0.730**	0.976	0.878	__0.710***__	0.734**
UNRATE	0.711*	0.690**	0.834	0.850	0.699**	0.792	0.996	0.735*	__0.679**__
GDPDEFL	0.849	0.866	0.962	0.870	0.843	__0.842__	0.947	0.854	0.851
GS10	0.874	__0.819**__	0.938	0.968	0.837*	0.825**	0.894	0.830**	0.824**
					$h = 4$				
PAYEMS	0.740	0.644**	0.836	0.894	0.638**	0.732	0.893	0.691**	__0.585**__
CPIAUCSL	1.013	__1.005__	1.015	1.008	1.036	1.033	1.053	1.031	1.027
FEDFUNDS	0.475***	0.476***	0.616***	0.577***	0.455***	0.457***	0.624**	__0.449***__	0.462***
GDP	1.031	0.920	0.885	0.958	0.863	0.998	0.945	__0.773**__	0.807**
UNRATE	0.705*	0.712**	0.864	0.918	0.698*	0.716	1.069	0.776*	__0.692**__
GDPDEFL	0.853**	0.849***	0.894	0.854*	0.864**	__0.841**__	0.943	0.872*	0.857**
GS10	0.896	0.901	0.981	0.961	0.911	0.893	0.972	0.896	__0.892__

This table reports the ratio between the MSFE of model i and the MSFE of the benchmark VAR(p) for the medium size VAR, computed as

$$MSFE_{ijh} = \frac{\sum_{\tau=\underline{t}}^{\bar{t}-h} e^2_{i,j,\tau+h}}{\sum_{\tau=\underline{t}}^{\bar{t}-h} e^2_{bcmk,j,\tau+h}},$$

where $p = 5$, $e^2_{i,j,\tau+h}$ and $e^2_{bcmk,j,\tau+h}$ are the squared forecast errors of variable j at time τ and forecast horizon h generated by model i and the VAR(p) model, respectively. $\underline{t}$ and $\bar{t}$ denote the start and end of the out-of-sample period, $i \in$ {DFM, FAVAR, BVAR-BGR, BVAR-GLP, BVAR-IP, SSVS, N-J, SNS, N-G}, $j \in$ {PAYEMS, CPIAUCSL,FEDFUNDS, GDP, UNRATE, GDPDEFL, GS10}, and $h = 1, ..., 4$. All forecasts are generated out-of-sample using recursive estimates of the models, with the out of sample period starting in 1985:Q1 and ending in 2015:Q4. Bold numbers indicate the lowest MSFE across all models for a given variable-forecast horizon pair. * significance at the 10% level; ** significance at the 5% level; *** significance at the 1% level.

Table 4. Out-of-sample point forecast performance, Large VAR

Variable	*DFM*	*FAVAR*	*BVAR-BGR*	*BVAR-GLP*	*BVAR-IP*	*SSVS*	*N-J*	*SNS*	*N-G*
$h = 1$									
PAYEMS	1.032	0.536**	0.585**	0.531**	0.541**	0.709*	0.493**	**0.463****	0.494**
CPIAUCSL	1.394	1.070	0.975	0.935	0.997	1.221	**0.927**	0.961	0.956
FEDFUNDS	0.390**	0.471***	0.649**	0.606**	**0.269****	0.313**	0.320**	0.341**	0.277**
GDP	1.169	0.834	0.792	0.802	0.713**	1.072	0.779*	**0.693*****	0.744**
UNRATE	0.659	0.627	0.689	0.801	0.648*	0.692**	0.692	0.632**	**0.627***
GDPDEFL	0.929	0.866	0.914	1.380	0.817	0.808	**0.742****	0.777*	0.767**
GS10	0.700**	0.672***	0.880	0.716***	**0.659*****	0.668***	0.701***	0.719**	0.666***
$h = 2$									
PAYEMS	0.817	0.520**	0.681*	0.628**	0.546**	0.681	0.534**	**0.451*****	0.465***
CPIAUCSL	1.080	1.041	0.957	1.001	**0.932**	0.984	0.971	0.952	0.986
FEDFUNDS	0.460***	0.421***	0.805*	0.917	0.390***	0.389***	0.424***	0.375***	**0.368*****
GDP	1.208	0.891	0.901	0.792**	0.720**	1.075	0.827	**0.695*****	0.739**
UNRATE	0.715*	**0.631****	0.743	0.807	0.699**	0.745*	0.783	0.673**	0.659**
GDPDEFL	0.805*	0.912	0.910	1.090	**0.801****	0.802*	0.824**	0.814**	0.829
GS10	0.884	0.859	1.053	1.010	0.803*	**0.786****	0.805*	0.796*	0.799*
$h = 3$									
PAYEMS	0.794	0.600**	0.736	0.710*	0.618**	0.680	0.640**	**0.530*****	0.534***
CPIAUCSL	0.954	0.948	0.995	1.061	0.944	**0.937**	1.012	0.966	0.968
FEDFUNDS	0.566***	0.549***	0.720***	0.763*	0.507***	0.532***	**0.506*****	0.510***	0.517***
GDP	1.011	0.884	0.848*	0.821*	0.780**	0.945	0.832*	**0.688*****	0.735**
UNRATE	0.697*	**0.611****	0.753	0.800*	0.698**	0.708**	0.796	0.673**	0.657**
GDPDEFL	0.862	0.863	0.906	1.074	0.861	**0.847**	0.888	0.852	0.866
GS10	0.849*	0.824**	0.937	0.941	0.841	0.824**	0.837*	0.827**	**0.823****
$h = 4$									
PAYEMS	0.749	0.654*	0.827	0.843	0.665*	0.676*	0.699*	0.585**	**0.578****
CPIAUCSL	1.016	**1.014**	1.025	1.035	1.031	1.022	1.050	1.040	1.023
FEDFUNDS	0.494***	0.477***	0.614***	0.553***	0.468***	0.471***	**0.458*****	0.463***	0.477***
GDP	1.052	1.015	0.878	0.905	0.867	1.007	0.907	**0.761*****	0.822*
UNRATE	0.726	**0.659****	0.876	0.950	0.718	0.704*	0.817	0.683*	0.685*
GDPDEFL	0.836**	**0.829*****	0.875*	0.935	0.843**	0.838**	0.867**	0.850**	0.834**
GS10	0.918	**0.901**	0.974	0.988	0.938	0.903	0.925	0.906	0.901

This table reports the ratio between the MSFE of model i and the MSFE of the benchmark VAR(p) for the large size VAR, computed as

$$MSFE_{ijh} = \frac{\sum_{\tau=\underline{t}}^{\bar{t}-h} e_{i,j,\tau+h}^{2}}{\sum_{\tau=\underline{t}}^{\bar{t}-h} e_{bcmk,j,\tau+h}^{2}},$$

where $p = 5$, $e_{i,j,\tau+h}^{2}$ and $e_{bcmk,j,\tau+h}^{2}$ are the squared forecast errors of variable j at time τ and forecast horizon h generated by model i and the VAR(p) model, respectively. $\underline{t}$ and $\bar{t}$ denote the start and end of the out-of-sample period, $i \in$ {DFM, FAVAR, BVAR-BGR, BVAR-GLP, BVAR-IP, SSVS, N-J, SNS, N-G}, $j \in$ {PAYEMS, CPIAUCSL,FEDFUNDS, GDP, UNRATE, GDPDEFL, GS10}, and $h = 1, ..., 4$. All forecasts are generated out-of-sample using recursive estimates of the models, with the out of sample period starting in 1985:Q1 and ending in 2015:Q4. Bold numbers indicate the lowest MSFE across all models for a given variable-forecast horizon pair. * significance at the 10% level; ** significance at the 5% level; *** significance at the 1% level.

Table 5. Out-of-sample point forecast performance, X-large VAR

Variable	*DFM*	*FAVAR*	*BVAR-BGR*	*BVAR-GLP*	*N-J*	*SNS*	*N-G*
				$h = 1$			
PAYEMS	0.456**	0.543**	0.473**	0.550**	0.574**	**0.445****	0.496**
CPIAUCSL	1.482	1.019	**0.902**	2.037	0.945	0.958	0.953
FEDFUNDS	0.725	0.875	0.577***	**0.220*****	0.322**	0.351**	0.283**
GDP	**0.590*****	0.674**	0.593***	0.818	0.775**	0.781*	0.749**
UNRATE	**0.474***	0.567*	0.602*	0.690	0.708*	0.642**	0.631*
GDPDEFL	1.038	0.957	1.039	2.381	**0.757****	0.788*	0.803*
GS10	0.726**	0.757*	0.853	0.769*	0.686***	0.701***	**0.664*****
				$h = 2$			
PAYEMS	0.491**	0.584**	0.539**	0.552**	0.654*	**0.444*****	0.476***
CPIAUCSL	1.123	1.035	0.978	1.550	0.973	**0.957**	0.962
FEDFUNDS	0.718**	0.697	0.733**	0.389***	0.407***	0.383***	**0.370*****
GDP	**0.615*****	0.728***	0.708***	0.698**	0.780*	0.800*	0.744**
UNRATE	**0.559*****	0.672**	0.696**	0.717*	0.803	0.668**	0.664**
GDPDEFL	0.935	0.869	0.885	1.505	0.819**	**0.785****	0.802*
GS10	0.994	0.987	1.129	1.167	0.832	**0.812***	0.830
				$h = 3$			
PAYEMS	0.577**	0.705	0.589**	0.612**	0.715	**0.497*****	0.534***
CPIAUCSL	0.969	**0.964**	1.008	1.174	1.002	0.985	0.973
FEDFUNDS	0.631***	0.528***	0.612***	0.591***	0.504***	**0.493*****	0.521***
GDP	**0.664*****	0.775***	0.729***	0.747**	0.788**	0.792**	0.738**
UNRATE	**0.616****	0.688**	0.660**	0.664**	0.817	0.635**	0.657**
GDPDEFL	0.887	0.866	0.924	1.403	0.861	0.874	**0.845**
GS10	0.899	0.848*	1.012	1.005	0.845*	0.836*	**0.828****
				$h = 4$			
PAYEMS	0.684**	0.800	0.643**	0.645**	0.857	**0.550****	0.580**
CPIAUCSL	1.037	1.029	1.059	1.561	1.070	1.038	**1.028**
FEDFUNDS	0.501***	0.492***	0.515***	0.595***	0.489***	**0.469*****	0.484***
GDP	**0.725*****	0.833*	0.734***	0.772***	0.869	0.871	0.833*
UNRATE	0.702*	0.736*	0.751*	0.696*	0.784	**0.643****	0.661*
GDPDEFL	0.880*	**0.845****	0.948	1.206	0.882*	0.858**	0.845**
GS10	0.971	0.950	1.104	1.171	0.929	0.930	**0.922**

This table reports the ratio between the MSFE of model i and the MSFE of the benchmark VAR(p) for the X-large size VAR, computed as

$$MSFE_{ijh} = \frac{\sum_{\tau=\underline{t}}^{\bar{t}-h} e^2_{i,j,\tau+h}}{\sum_{\tau=\underline{t}}^{\bar{t}-h} e^2_{bcmk,j,\tau+h}},$$

where $p = 5$, $e^2_{i,j,\tau+h}$ and $e^2_{bcmk,j,\tau+h}$ are the squared forecast errors of variable j at time τ and forecast horizon h generated by model i and the VAR(p) model, respectively. $\underline{t}$ and $\bar{t}$ denote the start and end of the out-of-sample period, $i \in$ {DFM, FAVAR, BVAR-BGR, BVAR-GLP, SSVS, N-J, SNS, N-G}, $j \in$ {PAYEMS, CPIAUCSL,FEDFUNDS, GDP, UNRATE, GDPDEFL, GS10}, and $h = 1, ..., 4$. All forecasts are generated out-of-sample using recursive estimates of the models, with the out of sample period starting in 1985:Q1 and ending in 2015:Q4. Bold numbers indicate the lowest MSFE across all models for a given variable-forecast horizon pair. * significance at the 10% level; ** significance at the 5% level; *** significance at the 1% level.

Table 6. Out-of-sample density forecast performance, Medium VAR

Variable	*DFM*	*FAVAR*	*BVAR-BGR*	*BVAR-GLP*	*BVAR-IP*	*SSVS*	*N-J*	*SNS*	*N-G*
					$h = 1$				
PAYEMS	-0.022	0.292	0.270*	0.144	0.336	0.119	0.396	**0.400**	0.355
CPIAUCSL	3.790	3.173	1.080	2.384	3.735	**4.721**	3.331	3.333	3.609
FEDFUNDS	0.598	0.576	0.411	0.281	0.582	0.516	**0.677**	0.595	0.583
GDP	0.051	0.117	0.188	**0.240**	0.200	-0.144	0.240	0.201	0.150
UNRATE	0.594	0.620	0.298	0.234	0.658	0.198	0.708	**0.745**	0.714
GDPDEFL	-0.097	-0.021	0.032	**0.054**	-0.070	-0.094	0.042	-0.005	-0.005
GS10	0.278*	0.301*	0.237*	0.327**	0.340**	0.317**	**0.353****	0.326**	0.339**
					$h = 2$				
PAYEMS	0.332	0.531	0.347	0.156	0.535	0.365	0.511	**0.595***	0.591*
CPIAUCSL	**1.252**	0.955	-0.906	-0.946	0.902	0.857	0.363	0.394	1.047
FEDFUNDS	0.053	0.082	**0.119*****	0.065	0.046	0.030	0.100**	0.109	0.053
GDP	-0.126	0.006	-0.021	-0.122	0.023	-0.101	0.065	**0.116**	0.069
UNRATE	0.412	0.424	0.069	0.064	0.417	0.296	0.370	0.479	**0.518**
GDPDEFL	0.010	-0.021	0.024	**0.078****	-0.018	-0.006	0.015	0.015	-0.014
GS10	0.064	0.112	0.016	0.035	0.092	0.105	0.086	**0.137***	0.134*
					$h = 3$				
PAYEMS	0.340	0.426*	0.026	-0.006	**0.517**	0.155	0.387	0.508	0.510
CPIAUCSL	1.421	1.355	-0.886	-0.288	0.878	1.368	0.801	0.513	**1.478**
FEDFUNDS	-0.054	-0.023	**0.102*****	0.066**	-0.052	-0.052	0.002	0.012	-0.038
GDP	-0.055	0.085	-0.074	-0.038	0.096	0.017	0.123	**0.209****	0.161*
UNRATE	0.470	0.410	0.107	0.210	0.362	0.349	0.450	0.372	**0.522**
GDPDEFL	-0.011	-0.001	0.021	**0.053**	-0.011	-0.005	0.011	0.005	-0.006
GS10	0.054	0.082	0.056	0.028	0.055	0.077	0.064	0.084	**0.091**
					$h = 4$				
PAYEMS	0.244	0.328	-0.428	-0.595	0.347	-0.067	0.246	0.356	**0.374**
CPIAUCSL	**1.013**	0.241	-0.845	-0.965	-0.012	0.589	0.456	0.288	0.345
FEDFUNDS	0.014	0.020	**0.158*****	0.134***	0.009	-0.005	0.042	0.066	0.012
GDP	-0.008	0.075	-0.292	-0.033	0.078	0.042	0.036	**0.199****	0.141*
UNRATE	0.651	**0.732**	-0.062	0.201	0.693	0.598	0.574	0.713	0.683
GDPDEFL	-0.004	0.017	0.067**	**0.086*****	-0.017	0.008	0.021	0.027	-0.001
GS10	0.003	0.022	0.007	0.029	-0.020	0.021	0.019	0.038	**0.043**

This table reports the average log predictive likelihood (ALPL) differential between model i and the benchmark VAR(p) for the medium VAR, computed as

$$ALPL_{ijh} = \frac{1}{\bar{t} - \underline{t} - h + 1} \sum_{\tau=\underline{t}}^{\bar{t}-h} \left(LPL_{i,j,\tau+h} - LPL_{bcmk,j,\tau+h} \right),$$

where $p = 5$, while $LPL_{i,j,\tau+h}$ and $LPL_{bcmk,j,\tau+h}$ are the log predictive likelihoods of variable j at time τ and forecast horizon h generated by model i and the VAR(p), respectively. $\underline{t}$ and $\bar{t}$ denote the start and end of the out-of-sample period, $i \in$ {DFM, BVAR-BGR, BVAR-GLP, BVAR-IP, SSVS, N-J, SNS, N-G}, $j \in$ {PAYEMS, CPIAUCSL,FEDFUNDS, GDP, UNRATE, GDPDEFL, GS10}, and $h = 1, ..., 4$. All forecasts are generated out-of-sample using recursive estimates of the models, with the out of sample period starting in 1985:Q1 and ending in 2015:Q4. Bold numbers indicate the highest ALPL across all models for a given variable-forecast horizon pair. * significance at the 10% level; ** significance at the 5% level; *** significance at the 1% level.

Table 7. Out-of-sample density forecast performance, Large VAR

Variable	*DFM*	*FAVAR*	*BVAR-BGR*	*BVAR-GLP*	*BVAR-IP*	*SSVS*	*N-J*	*SNS*	*N-G*
					$h = 1$				
PAYEMS	0.526	0.856	0.759	0.830	0.770	0.658	**0.897**	0.844	0.790
CPIAUCSL	1.654	1.935	-0.031	-0.269	2.465	**3.216**	2.113*	2.565	2.439
FEDFUNDS	0.317	0.295	-0.107	0.044	0.251	0.225	**0.327**	0.262	0.269
GDP	-0.072	0.082	**0.195***	0.143	0.085	-0.035	0.137	0.150*	0.139
UNRATE	1.016	**1.093**	0.926	0.900	1.017	0.991	0.991	0.977	1.011
GDPDEFL	-0.065	-0.006	**0.045**	-0.169	-0.050	-0.010	0.038	0.007	0.006
GS10	0.298**	**0.325****	0.133	0.286**	0.275*	0.307**	0.312**	0.299**	0.317**
					$h = 2$				
PAYEMS	0.268	0.497*	0.124	0.382*	0.421	0.267	**0.520***	0.495*	0.456
CPIAUCSL	0.588	0.758	-0.716	-1.438	1.037	**1.506**	-0.063	1.246	0.750
FEDFUNDS	0.058	0.077	0.076	-0.277	0.025	0.031	**0.104**	0.069	0.045
GDP	-0.061	0.141	0.089	0.041	0.147	-0.010	0.190	**0.240****	0.091
UNRATE	0.303	**0.414**	-0.070	0.123	0.281	0.218	0.214*	0.366*	0.354
GDPDEFL	0.025	-0.002	**0.057**	-0.133	-0.004	0.023	0.034	0.032	0.009
GS10	0.069	0.082	-0.052	-0.031	0.079	0.105	0.127*	0.116	**0.141***
					$h = 3$				
PAYEMS	0.191	0.258*	-0.204	-0.327	0.303*	0.231*	0.273*	**0.342***	0.335
CPIAUCSL	**2.944**	2.522	-0.056	-0.946	1.612	2.810	1.139	1.629	2.032
FEDFUNDS	-0.017	-0.018	**0.158*****	0.085**	-0.071	-0.034	0.046	0.010	-0.020
GDP	0.001	0.087	-0.045	-0.003	0.158	0.078	0.114*	**0.197*****	0.179**
UNRATE	**0.369**	0.291	0.107	-0.017	0.308	0.332	0.047	0.260**	0.147
GDPDEFL	-0.004	-0.012	**0.064**	-0.208	-0.035	0.002	0.003	0.008	-0.008
GS10	0.081	0.090	0.063	0.040	0.056	0.088	0.099**	0.105*	**0.110***
					$h = 4$				
PAYEMS	0.359	0.341	-0.277	-0.525	0.345	0.282**	0.380*	**0.471**	0.389
CPIAUCSL	1.766	1.320	-0.231	-1.023	1.015	**2.380**	0.850	0.924	1.496
FEDFUNDS	-0.005	0.020	**0.193*****	0.134***	-0.027	-0.003	0.083**	0.043	0.019
GDP	-0.017	-0.011	-0.130	-0.100	0.012	-0.042	-0.018	**0.121****	-0.028
UNRATE	0.483	0.452	-0.281	-0.209	0.473	**0.517**	-0.038	0.281*	0.292
GDPDEFL	0.018	0.003	**0.071****	-0.196	-0.018	0.006	0.030	0.011	0.002
GS10	-0.004	0.003	0.021	-0.006	-0.038	0.001	**0.026**	0.025	0.026

This table reports the average log predictive likelihood (ALPL) differential between model i and the benchmark VAR(p) for the large VAR, computed as

$$ALPL_{ijh} = \frac{1}{\bar{t} - \underline{t} - h + 1} \sum_{\tau=\underline{t}}^{\bar{t}-h} \left(LPL_{i,j,\tau+h} - LPL_{bcmk,j,\tau+h} \right),$$

where $p = 5$, while $LPL_{i,j,\tau+h}$ and $LPL_{bcmk,j,\tau+h}$ are the log predictive likelihoods of variable j at time τ and forecast horizon h generated by model i and the VAR(p), respectively. $\underline{t}$ and $\bar{t}$ denote the start and end of the out-of-sample period, $i \in$ {DFM, FAVAR, BVAR-BGR, BVAR-GLP, BVAR-IP, SSVS, N-J, SNS, N-G}, $j \in$ {PAYEMS, CPIAUCSL,FEDFUNDS, GDP, UNRATE, GDPDEFL, GS10}, and $h = 1, ..., 4$. All forecasts are generated out-of-sample using recursive estimates of the models, with the out of sample period starting in 1985:Q1 and ending in 2015:Q4. Bold numbers indicate the highest ALPL across all models for a given variable-forecast horizon pair. * significance at the 10% level; ** significance at the 5% level; *** significance at the 1% level.

Table 8. Out-of-sample density forecast performance, X-large VAR

Variable	*DFM*	*FAVAR*	*BVAR-BGR*	*BVAR-GLP*	*N-J*	*SNS*	*N-G*
				$h = 1$			
PAYEMS	0.573	0.571	**0.641***	0.530*	0.555	0.615	0.531
CPIAUCSL	0.164	0.539	-4.509	-2.116	0.269	**0.692***	0.105
FEDFUNDS	0.232	-0.352	-0.669	**0.782***	0.503	0.442	0.450
GDP	0.273**	0.233**	**0.282*****	0.069	0.054	0.173*	0.061
UNRATE	**0.890**	0.843	0.692	0.547	0.627	0.726	0.750
GDPDEFL	-0.053	-0.012	-0.021	-0.526	**0.059**	0.040	0.020
GS10	0.181*	0.108	-0.405	0.086	0.203***	0.187**	**0.215*****
				$h = 2$			
PAYEMS	0.272**	0.196*	-0.255	-0.909	0.095	**0.292*****	0.227**
CPIAUCSL	-0.251	**0.450**	-3.534	-0.124	-0.355	-0.883	-0.862
FEDFUNDS	0.048	-0.009	-0.184	**0.264*****	0.123	0.129*	0.107
GDP	**0.159*****	0.024	-0.010	0.010	0.017	0.092*	0.038
UNRATE	**0.417****	0.225**	-0.627	0.244**	0.068	0.226***	0.408*
GDPDEFL	0.034	0.053	**0.093**	-0.240	0.061**	0.080***	0.070**
GS10	0.000	0.006	-0.536	-0.116	**0.101***	0.084	0.090
				$h = 3$			
PAYEMS	0.254***	0.139	-0.599	-0.932	-0.017	**0.284****	0.250*
CPIAUCSL	-0.559	0.367	-3.237	**1.070**	-0.927	-0.582	-0.157
FEDFUNDS	0.054**	0.071***	**0.353*****	0.164***	0.082**	0.098***	0.058**
GDP	**0.196*****	0.137***	-0.260	0.099	0.067	0.099*	0.144***
UNRATE	0.193**	0.240**	-1.018	0.052	-0.186	0.134	**0.256*****
GDPDEFL	0.051**	0.054*	**0.117**	-0.314	0.063**	0.059**	0.050
GS10	0.038	0.067**	-0.053	0.007	**0.094****	0.076**	0.073**
				$h = 4$			
PAYEMS	0.091	0.107	-1.145	-1.217	-0.250	**0.222****	0.193*
CPIAUCSL	0.463	0.820	-2.714	**1.267**	0.273	0.264	0.227
FEDFUNDS	0.088***	0.097***	**0.401*****	0.163***	0.108***	0.133***	0.093***
GDP	**0.165*****	0.047	0.034	0.041	-0.127	0.002	0.031
UNRATE	0.025	-0.046	-1.277	-0.234	-0.327	-0.131	**0.190****
GDPDEFL	0.056***	0.069***	**0.083**	-0.342	0.056**	0.075***	0.073***
GS10	0.011	0.025	-0.049	-0.050	**0.066***	0.046	0.055**

This table reports the average log predictive likelihood (ALPL) differential between model i and the benchmark VAR(p) for the XX-large VAR, computed as

$$ALPL_{ijh} = \frac{1}{\bar{t} - \underline{t} - h + 1} \sum_{\tau=\underline{t}}^{\bar{t}-h} \left(LPL_{i,j,\tau+h} - LPL_{bcmk,j,\tau+h} \right),$$

where $p = 5$, while $LPL_{i,j,\tau+h}$ and $LPL_{bcmk,j,\tau+h}$ are the log predictive likelihoods of variable j at time τ and forecast horizon h generated by model i and the VAR(p), respectively. $\underline{t}$ and $\bar{t}$ denote the start and end of the out-of-sample period, $i \in$ {DFM, FAVAR, BVAR-BGR, BVAR-GLP, N-J, SNS, N-G}, $j \in$ {PAYEMS, CPIAUCSL,FEDFUNDS, GDP, UNRATE, GDPDEFL, GS10}, and $h = 1, ..., 4$. All forecasts are generated out-of-sample using recursive estimates of the models, with the out of sample period starting in 1985:Q1 and ending in 2015:Q4. Bold numbers indicate the highest ALPL across all models for a given variable-forecast horizon pair. * significance at the 10% level; ** significance at the 5% level; *** significance at the 1% level.

Appendix A Technical appendix

In this section, we provide detailed derivations and proofs of all the main results in the paper.

A.1 Derivation of the rotated regression and rotated likelihood

We begin by providing details on the derivation of the rotated regression in equation (3) and the joint likelihood of the rotated data in (4). Start with the simple univariate linear regression model in (1), which for convenience we report here

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{v}, \tag{A.1}$$

Next, introduce the $T \times T$ full-rank rotation matrix $\boldsymbol{Q}_j = \left[\boldsymbol{q}_j, \boldsymbol{W}_j\right]$ where $\boldsymbol{q}_j = \boldsymbol{X}_j / \left\|\boldsymbol{X}_j\right\|$ and $\boldsymbol{W}_j$ is an arbitrarily chosen $T \times (T-1)$ matrix subject to the constraint $\boldsymbol{W}_j\boldsymbol{W}_j' = \boldsymbol{I}_T - \boldsymbol{q}_j\boldsymbol{q}_j'$. Next, rewrite (A.1) as

$$\boldsymbol{y} = \boldsymbol{X}_j\beta_j + \boldsymbol{X}_{(-j)}\boldsymbol{\beta}_{(-j)} + \boldsymbol{v} \tag{A.2}$$

where $\boldsymbol{X}_{(-j)} = \boldsymbol{X} \setminus \boldsymbol{X}_j$ and $\boldsymbol{\beta}_{(-j)} = \boldsymbol{\beta} \setminus \beta_j$. Proceed by pre-multiplying both LHS and RHS of (A.2) by $\boldsymbol{Q}_j'$, to obtain

$$\boldsymbol{Q}_j'\boldsymbol{y} = \boldsymbol{Q}_j'\boldsymbol{X}_j\beta_j + \boldsymbol{Q}_j'\boldsymbol{X}_{(-j)}\boldsymbol{\beta}_{(-j)} + \boldsymbol{Q}_j'\boldsymbol{v}, \tag{A.3}$$

or, using the fact that $\boldsymbol{Q}_j = \left[\boldsymbol{q}_j, \boldsymbol{W}_j\right]$,

$$\begin{bmatrix} \boldsymbol{q}_j' \\ \boldsymbol{W}_j' \end{bmatrix} \boldsymbol{y} = \begin{bmatrix} \boldsymbol{q}_j' \\ \boldsymbol{W}_j' \end{bmatrix} \boldsymbol{X}_j\beta_j + \begin{bmatrix} \boldsymbol{q}_j' \\ \boldsymbol{W}_j' \end{bmatrix} \boldsymbol{X}_{(-j)}\boldsymbol{\beta}_{(-j)} + \begin{bmatrix} \boldsymbol{q}_j' \\ \boldsymbol{W}_j' \end{bmatrix} \boldsymbol{v}. \tag{A.4}$$

Now using the definition of $\boldsymbol{q}_j$ and the formulas for y_j^* and $\widetilde{\boldsymbol{y}}_j$ in (2), we have that

$$\begin{bmatrix} y_j^* \\ \widetilde{\boldsymbol{y}}_j \end{bmatrix} = \begin{bmatrix} \left(\boldsymbol{X}_j'\boldsymbol{X}_j / \left\|\boldsymbol{X}_j\right\|\right) \\ \boldsymbol{W}_j'\boldsymbol{q}_j \left\|\boldsymbol{X}_j\right\| \end{bmatrix} \beta_j + \begin{bmatrix} \boldsymbol{q}_j'\boldsymbol{X}_{(-j)} \\ \boldsymbol{W}_j'\boldsymbol{X}_{(-j)} \end{bmatrix} \boldsymbol{\beta}_{(-j)} + \begin{bmatrix} \boldsymbol{q}_j'\boldsymbol{v} \\ \boldsymbol{W}_j'\boldsymbol{v} \end{bmatrix}, \tag{A.5}$$

Further simplifications lead to (3), i.e.

$$\begin{bmatrix} y_j^* \\ \widetilde{\boldsymbol{y}}_j \end{bmatrix} = \begin{bmatrix} \left\|\boldsymbol{X}_j\right\| \beta_j \\ \boldsymbol{0} \end{bmatrix} + \begin{bmatrix} \boldsymbol{X}_{(-j)}^*\boldsymbol{\beta}_{(-j)} \\ \widetilde{\boldsymbol{X}}_{(-j)}\boldsymbol{\beta}_{(-j)} \end{bmatrix} + \begin{bmatrix} v_j^* \\ \widetilde{\boldsymbol{v}}_j \end{bmatrix}, \tag{A.6}$$

where we have exploited the following two results:

1. $\left(\boldsymbol{X}_j'\boldsymbol{X}_j / \left\|\boldsymbol{X}_j\right\|\right) = \left\|\boldsymbol{X}_j\right\|$. This is due to the fact that $\boldsymbol{X}_j'\boldsymbol{X}_j = \left\|\boldsymbol{X}_j\right\|^2$;

2. By definition, $\boldsymbol{W}_j$ and $\boldsymbol{q}_j$ are orthogonal. They all are columns of the orthogonal matrix $\boldsymbol{Q}_j$, so by construction $\boldsymbol{W}_j'\boldsymbol{q}_j = \mathbf{0}$.

Next, to go from (3) to (4), note that $E\left(\boldsymbol{Q}_j'\boldsymbol{v}\right) = \mathbf{0}$ while $var\left(\boldsymbol{Q}_j'\boldsymbol{v}\right) = \sigma^2\boldsymbol{Q}_j'\boldsymbol{Q}_j = \sigma^2\boldsymbol{I}_T$ which, combined with (A.6), leads to the rotated likelihood in equation (4). ■

A.2 Links with partitioned regression method

There are a number of similarities between the rotation we introduced in Section 2 and the traditional partitioned regression (or "partial-time regression", using the terminology of Frisch and Waugh, 1933). Suppose, along the lines of our discussion in Section 2, that we are interested in β_j $(j = 1, ..., k)$, the j-th element of a vector of coefficients $\boldsymbol{\beta}$ in a standard homoskedastic multivariate regression model. The standard partitioned regression method works by first defining the $T \times T$ matrix $\boldsymbol{M}_j = \boldsymbol{I}_T - \boldsymbol{X}_j\left(\boldsymbol{X}_j'\boldsymbol{X}_j\right)^{-1}\boldsymbol{X}_j'$. It is easy to show using the algebra of partitioned matrices that $\widehat{\beta}_j$, the OLS estimates of β_j can be obtained as the solution of

$$\widehat{\beta}_j = \left(\boldsymbol{X}_j'\boldsymbol{X}_j\right)^{-1}\boldsymbol{X}_j'\left(\boldsymbol{y} - \boldsymbol{X}_{(-j)}\widehat{\boldsymbol{\beta}}_{(-j)}\right) \tag{A.7}$$

where the sub-vector $\widehat{\boldsymbol{\beta}}_{(-j)}$ is the solution of the following regression

$$\widehat{\boldsymbol{\beta}}_{(-j)} = \left(\boldsymbol{X}_{(-j)}^{\dagger\prime}\boldsymbol{X}_{(-j)}^{\dagger}\right)^{-1}\boldsymbol{X}_{(-j)}^{\dagger\prime}\boldsymbol{y}^{\dagger} \tag{A.8}$$

with $\boldsymbol{X}_{(-j)}^{\dagger} = \boldsymbol{M}_j\boldsymbol{X}_{(-j)}$ and $\boldsymbol{y}^{\dagger} = \boldsymbol{M}_j\boldsymbol{y}$ denoting the projections of $\boldsymbol{X}_{(-j)}$ and $\boldsymbol{y}$ on a space that is orthogonal to $\boldsymbol{X}_j$.

The two-step approach behind the partitioned regression method in (A.7)-(A.8) is very closely related to the two-step procedure implied by our proposed approach. However, there are some important differences:

- The rotation implied the $\boldsymbol{Q}_j$ matrix we rely on preserves homoskedasticity in the rotated regression. In contrast, the rotation implied by the annihilator matrix $\boldsymbol{M}_j$ in the partitioned regression method transforms the original homoskedastic regression into a heteroskedastic model in the rotated space. This can be easily seen by noting that $\widehat{\boldsymbol{\beta}}_{(-j)}$

in (A.8) is the OLS solution to the following regression model[36]

$$\boldsymbol{y}^{\dagger} = \boldsymbol{X}^{\dagger}_{(-j)}\boldsymbol{\beta}_{(-j)} + \boldsymbol{v}^{\dagger}_{j} \qquad \boldsymbol{v}^{\dagger}_{j} \sim \mathcal{N}\left(\mathbf{0}, \sigma^2 \boldsymbol{M}_j\right) \tag{A.9}$$

- The rotation implied by the matrix $\boldsymbol{Q}_j$ we introduce allows to split the $T \times 1$ vector $\boldsymbol{y}$ into the scalar y^*_j, which does depend on β_j, and the remaining $T-1$ observations $\widetilde{\boldsymbol{y}}_j$, which do not depend on β_j. Combined with the previous point (i.e., the homoskedasticity-preserving rotation), this is what allows us to estimate $\widehat{\boldsymbol{\beta}}_{(-j)}$ using $T-1$ observations and $\widehat{\beta}_j$ using a single observation. This, in turn, leads to the expression for the marginal posterior of β_j in equation (5), the expression for the rotated marginal likelihood in equation (6) and, as a byproduct, the quality of its approximation and the low computational costs needed to implement adaptive hierarchical priors in this setting.

A.3 Derivation of the rotated conditional likelihood

In this subsection, we provide details on the results in equations (6), (7), and (8). Start by focusing on the top row of (4), and note that the conditional density $p\left(y^*_j|\boldsymbol{\beta},\sigma^2\right)$ can be decomposed as follows

$$y^*_j = \|\boldsymbol{X}_j\|\,\beta_j + y^+_j \tag{A.10}$$

where

$$y^+_j|\boldsymbol{\beta}_{(-j)},\sigma^2 \sim \mathcal{N}\left(\boldsymbol{X}^*_{(-j)}\boldsymbol{\beta}_{(-j)},\sigma^2\right) \tag{A.11}$$

Notice that the newly defined $p\left(y^+_j|\boldsymbol{\beta}_{(-j)},\sigma^2\right)$ can be interpreted as essentially the predictive distribution associated with the auxiliary regression that is defined in the second row of (4). This leads to the following result,

$$\begin{aligned} p\left(y^*_j|\beta_j,\widetilde{\boldsymbol{y}}_j\right) &= \|\boldsymbol{X}_j\|\,\beta_j + p\left(y^+_j|\widetilde{\boldsymbol{y}}_j\right) \\ &= \|\boldsymbol{X}_j\|\,\beta_j + \int\int p\left(y^+_j|\boldsymbol{\beta}_{(-j)},\sigma^2,\widetilde{\boldsymbol{y}}_j\right)p\left(\boldsymbol{\beta}_{(-j)},\sigma^2|\widetilde{\boldsymbol{y}}_j\right)d\boldsymbol{\beta}_{(-j)}d\sigma^2 \end{aligned} \tag{A.12}$$

The key to solving (A.12) is to compute the integral in the second row of the equation, which in turn will depend on the prior distribution adopted for $p\left(\boldsymbol{\beta}_{(-j)},\sigma^2\right)$. As we discussed in

[36] This is due to the fact that $\boldsymbol{M}_j$ is both symmetric and idempotent, leading to $Var\left(\boldsymbol{v}^{\dagger}_j\right) = Var\left(\boldsymbol{M}'_j\boldsymbol{v}\right) = \sigma^2\boldsymbol{M}'_j\boldsymbol{M}_j = \sigma^2\boldsymbol{M}_j$.

Section 2, for computational tractability we chose to rely on the natural conjugate prior,

$$\begin{aligned} \boldsymbol{\beta}_{(-j)}|\sigma^2 &\sim \mathcal{N}\left(\underline{\boldsymbol{\beta}}_{(-j)}, \sigma^2 \underline{\boldsymbol{V}}_{\boldsymbol{\beta}_{(-j)}}\right) \\ \sigma^2 &\sim \mathcal{IG}\left(\underline{\psi}, \underline{d}\right) \end{aligned} \tag{A.13}$$

It is straightforward to show that the posterior distribution $p\left(\boldsymbol{\beta}_{(-j)}, \sigma^2|\widetilde{\boldsymbol{y}}_j\right)$ also belongs to the Normal-Inverse-Gamma (NIG) family, and is given by

$$\begin{aligned} \boldsymbol{\beta}_{(-j)}|\sigma^2, \widetilde{\boldsymbol{y}}_j &\sim \mathcal{N}\left(\overline{\boldsymbol{\beta}}_{(-j)}, \sigma^2 \overline{\boldsymbol{V}}_{\boldsymbol{\beta}_{(-j)}}\right) \\ \sigma^2|\widetilde{\boldsymbol{y}}_j &\sim \mathcal{IG}\left(\overline{\psi}_{(-j)}, \overline{d}\right) \end{aligned} \tag{A.14}$$

where $\overline{d} = \underline{d} + (T-1)/2$,

$$\overline{\boldsymbol{V}}_{\boldsymbol{\beta}_{(-j)}} = \left(\underline{\boldsymbol{V}}^{-1}_{\boldsymbol{\beta}_{(-j)}} + \widetilde{\boldsymbol{X}}'_{(-j)}\widetilde{\boldsymbol{X}}_{(-j)}\right)^{-1}, \tag{A.15}$$

$$\overline{\boldsymbol{\beta}}_{(-j)} = \overline{\boldsymbol{V}}_{\boldsymbol{\beta}_{(-j)}}\left(\underline{\boldsymbol{V}}^{-1}_{\boldsymbol{\beta}_{(-j)}}\underline{\boldsymbol{\beta}}_{(-j)} + \widetilde{\boldsymbol{X}}'_{(-j)}\widetilde{\boldsymbol{y}}_j\right), \tag{A.16}$$

and

$$\overline{\psi}_{(-j)} = \underline{\psi} + \frac{1}{2}\left(\widetilde{\boldsymbol{y}}'_j\widetilde{\boldsymbol{y}}_j + \underline{\boldsymbol{\beta}}'_{(-j)}\underline{\boldsymbol{V}}^{-1}_{\boldsymbol{\beta}_{(-j)}}\underline{\boldsymbol{\beta}}_{(-j)} - \overline{\boldsymbol{\beta}}'_{(-j)}\overline{\boldsymbol{V}}^{-1}_{\boldsymbol{\beta}_{(-j)}}\overline{\boldsymbol{\beta}}_{(-j)}\right). \tag{A.17}$$

Armed with an analytical expression for the posterior $p\left(\boldsymbol{\beta}_{(-j)}, \sigma^2|\widetilde{\boldsymbol{y}}_j\right)$, we are now ready to derive the rotated conditional likelihood:

$$\begin{aligned} p\left(y^*_j|\beta_j, \widetilde{\boldsymbol{y}}_j\right) &= \|\boldsymbol{X}_j\|\,\beta_j + \int\int p\left(y^+_j|\boldsymbol{\beta}_{(-j)}, \sigma^2, \widetilde{\boldsymbol{y}}_j\right) p\left(\boldsymbol{\beta}_{(-j)}, \sigma^2|\widetilde{\boldsymbol{y}}_j\right) d\boldsymbol{\beta}_{(-j)} d\sigma^2 \\ &= \|\boldsymbol{X}_j\|\,\beta_j + \int\int \mathcal{N}\left(\boldsymbol{X}^*_{(-j)}\boldsymbol{\beta}_{(-j)}, \sigma^2\right) \times \\ &\times \mathcal{N}\left(\overline{\boldsymbol{\beta}}_{(-j)}, \sigma^2\overline{\boldsymbol{V}}_{\boldsymbol{\beta}_{(-j)}}\right) \mathcal{IG}\left(\overline{\psi}_{(-j)}, \overline{d}\right) d\boldsymbol{\beta}_{(-j)} d\sigma^2 \\ &= \|\boldsymbol{X}_j\|\,\beta_j + t_{2\overline{d}}\left(\overline{\mu}_j, \overline{\tau}^2_j\right) \\ &\approx \|\boldsymbol{X}_j\|\,\beta_j + \mathcal{N}\left(\overline{\mu}_j, \overline{\tau}^2_j\right) \end{aligned} \tag{A.18}$$

where

$$\overline{\mu}_j = \boldsymbol{X}^*_{(-j)}\overline{\boldsymbol{\beta}}_{(-j)} \tag{A.19}$$

and

$$\overline{\tau}^2_j = \frac{\overline{\psi}_{(-j)}}{\overline{d}}\left(1 + \boldsymbol{X}^*_{(-j)}\overline{\boldsymbol{V}}_{\boldsymbol{\beta}_{(-j)}}\boldsymbol{X}^{*\prime}_{(-j)}\right). \tag{A.20}$$

This conclude the derivations of equations (6), (7), and (8). ■

A.4 Calculation of optimal shrinkage intensity under a Normal-Jeffreys prior

Start with the approximation in (6), which here we slightly rearrange to be

$$\left(y_j^* - \overline{\mu}_j\right)|\beta_j, \widetilde{\boldsymbol{y}}_j \sim \mathcal{N}\left(\left\|\boldsymbol{X}_j\right\|\beta_j, \overline{\tau}_j^2\right),$$

and write the Normal-Jeffreys prior as in (9)

$$\beta_j|\lambda_j^2 \sim \mathcal{N}\left(0, \lambda_j^2 \underline{V}_{\beta_j}\right) \tag{A.21}$$

Next, the marginal likelihood $p\left(y_j^* - \overline{\mu}_j \middle| \lambda_j^2, \widetilde{\boldsymbol{y}}_j\right)$ is given by

$$\begin{aligned} p\left(y_j^* - \overline{\mu}_j\right| \lambda_j^2, \widetilde{\boldsymbol{y}}_j\left.\right) &= \int p\left(y_j^* - \overline{\mu}_j\right| \beta_j, \widetilde{\boldsymbol{y}}_j\left.\right) p\left(\beta_j\right|\lambda_j^2\left.\right) d\beta_j \\ &= \mathcal{N}\left(y_j^* - \overline{\mu}_j \middle| \left\|\boldsymbol{X}_j\right\|^2 \lambda_j^2 \underline{V}_{\beta_j} + \overline{\tau}_j^2\right). \end{aligned} \tag{A.22}$$

or, more explicitly,

$$p\left(y_j^* - \overline{\mu}_j|\lambda_j^2, \widetilde{\boldsymbol{y}}_j\right) = \frac{1}{\sqrt{2\pi}\sqrt{\overline{\tau}_j^2 + \left\|\boldsymbol{X}_j\right\|^2 \lambda_j^2 \underline{V}_{\beta_j}}} \times \exp\left(-\frac{\left(y_j^* - \overline{\mu}_j\right)^2}{2\left(\overline{\tau}_j^2 + \left\|\boldsymbol{X}_j\right\|^2 \lambda_j^2 \underline{V}_{\beta_j}\right)}\right)$$

To find the λ_j^2 that maximizes $p\left(y_j^* - \overline{\mu}_j\right)|\lambda_j^2, \widetilde{\boldsymbol{y}}_j$, take the log and only focus on the terms that involve λ_j^2:

$$\ln p\left(y_j^* - \overline{\mu}_j|\lambda_j^2, \widetilde{\boldsymbol{y}}_j\right) \propto -\frac{1}{2}\ln\left(\overline{\tau}_j^2 + \left\|\boldsymbol{X}_j\right\|^2 \lambda_j^2 \underline{V}_{\beta_j}\right) - \frac{1}{2}\frac{\left(y_j^* - \overline{\mu}_j\right)^2}{\left(\overline{\tau}_j^2 + \left\|\boldsymbol{X}_j\right\|^2 \lambda_j^2 \underline{V}_{\beta_j}\right)}$$

Now taking the derivative with respect to λ_j^2 and setting it to zero

$$\frac{\partial \ln p\left(y_j^* - \overline{\mu}_j|\lambda_j^2, \widetilde{\boldsymbol{y}}_j\right)}{\partial \lambda_j^2} = -\frac{1}{2}\frac{\left\|\boldsymbol{X}_j\right\|^2 \underline{V}_{\beta_j}}{\left(\overline{\tau}_j^2 + \left\|\boldsymbol{X}_j\right\|^2 \lambda_j^2 \underline{V}_{\beta_j}\right)} + \frac{1}{2}\frac{\left(y_j^* - \overline{\mu}_j\right)^2 \left\|\boldsymbol{X}_j\right\|^2 \underline{V}_{\beta_j}}{\left(\overline{\tau}_j^2 + \left\|\boldsymbol{X}_j\right\|^2 \lambda_j^2 \underline{V}_{\beta_j}\right)^2} = 0$$

leads to the solution in (12),

$$\widehat{\lambda}_j^2 = \max\left[0, \frac{\left(y_j^* - \overline{\mu}_j\right)^2 - \overline{\tau}_j^2}{\left\|\boldsymbol{X}_j\right\|^2 \underline{V}_{\beta_j}}\right]. \tag{A.23}$$

■

A.5 Derivation of posterior probability of inclusion under a Spike-and-Slab prior

Start with (18), which for convenience we rewrite here as

$$\widehat{\pi}_j = p\left(\lambda_j = 1 \middle|\, y_j^*, \widetilde{\boldsymbol{y}}_j\right) = \frac{p\left(y_j^* \middle|\, \lambda_j = 1, \widetilde{\boldsymbol{y}}_j\right) p\left(\lambda_j = 1\right)}{p\left(y_j^* \middle| \lambda_j = 0, \widetilde{\boldsymbol{y}}_j\right) p\left(\lambda_j = 0\right) + p\left(y_j^* \middle| \lambda_j = 1, \widetilde{\boldsymbol{y}}_j\right) p\left(\lambda_j = 1\right)} \tag{A.24}$$

Next, notice that $= p\left(\lambda_j = 1\right) = \underline{\pi}_0$ and $p\left(\lambda_j = 0\right) = 1 - \underline{\pi}_0$. Furthermore, the approximation in (6) along with the independence between β_j and $\widetilde{\boldsymbol{y}}_j$ imply that

$$\begin{aligned} p\left(y_j^* \middle|\, \lambda_j = 1, \widetilde{\boldsymbol{y}}_j\right) &\approx \int p\left(y_j^* \middle|\, \beta_j, \lambda_j = 1, \widetilde{\boldsymbol{y}}_j\right) p\left(\beta_j \middle| \lambda_j = 1, \widetilde{\boldsymbol{y}}_j\right) d\beta_j \\ &\approx \int p\left(y_j^* \middle|\, \beta_j, \lambda_j = 1, \widetilde{\boldsymbol{y}}_j\right) p\left(\beta_j \middle| \lambda_j = 1\right) d\beta_j \\ &\sim \mathcal{N}\left(y_j^* \middle| \overline{\mu}_j, \overline{\tau}_j^2 + \left\|\boldsymbol{X}_j\right\|^2 \underline{V}_{\beta_j}\right) \end{aligned} \tag{A.25}$$

while, similarly,

$$\begin{aligned} p\left(y_j^* \middle|\, \lambda_j = 0, \widetilde{\boldsymbol{y}}_j\right) &\approx \int p\left(y_j^* \middle|\, \beta_j, \lambda_j = 0, \widetilde{\boldsymbol{y}}_j\right) p\left(\beta_j \middle| \lambda_j = 0, \widetilde{\boldsymbol{y}}_j\right) d\beta_j \\ &\approx \int p\left(y_j^* \middle|\, \beta_j, \lambda_j = 0, \widetilde{\boldsymbol{y}}_j\right) p\left(\beta_j \middle| \lambda_j = 0\right) d\beta_j \\ &\sim \mathcal{N}\left(y_j^* \middle| \overline{\mu}_j, \overline{\tau}_j^2\right) \end{aligned} \tag{A.26}$$

Plugging (A.25) and (A.26) into (A.24) leads to (19). ■

A.6 Triangularization of the VAR

Start from the n-dimensional VAR(p) model in (21), which for convenience we rewrite here

$$\boldsymbol{y}_t = \boldsymbol{c} + \boldsymbol{A}_1 \boldsymbol{y}_{t-1} + \ldots + \boldsymbol{A}_p \boldsymbol{y}_{t-p} + \boldsymbol{\varepsilon}_t, \quad t = 1, ..., T, \tag{A.27}$$

where $\boldsymbol{y}_t$ is an $n \times 1$ vector of time series of interest, $\boldsymbol{c}$ is an $n \times 1$ vector of intercepts, $\boldsymbol{A}_1, ..., \boldsymbol{A}_p$ are $n \times n$ matrices of coefficients on the lagged dependent variables, and $\boldsymbol{\varepsilon}_t \sim \mathcal{N}\left(\boldsymbol{0}, \boldsymbol{\Omega}\right)$, with $\boldsymbol{\Omega}$ an $n \times n$ covariance matrix. Next, following Carriero et al. (2017), decompose the VAR covariance matrix $\boldsymbol{\Omega}$ in (A.27) as $\boldsymbol{\Omega} = \boldsymbol{\Gamma}^{-1} \boldsymbol{\Sigma} \left(\boldsymbol{\Gamma}^{-1}\right)'$, where

$$\boldsymbol{\Gamma}^{-1} = \begin{bmatrix} 1 & 0 & \cdots & 0 & 0 \\ \gamma_{2,1} & 1 & \ddots & \vdots & \vdots \\ \vdots & \ddots & \ddots & 0 & 0 \\ \gamma_{n-1,1} & \cdots & \gamma_{n-1,n-2} & 1 & 0 \\ \gamma_{n,1} & \cdots & \gamma_{n,n-2} & \gamma_{n,n-1} & 1 \end{bmatrix}, \tag{A.28}$$

and $\boldsymbol{\Sigma} = diag\left(\sigma_1^2, ..., \sigma_n^2\right)$. Under this decomposition the residuals of the original VAR(p) in (A.27) can be written using the identity $\boldsymbol{\varepsilon}_t = \boldsymbol{\Gamma}^{-1}\boldsymbol{\Sigma}^{1/2}\boldsymbol{u}_t$, with $\boldsymbol{u}_t \sim \mathcal{N}\left(\mathbf{0}, \boldsymbol{I}_n\right)$, which implies that the i-th row of this identity is

$$\varepsilon_{i,t} = \gamma_{i,1}\sigma_1 u_{1,t} + ... + \gamma_{i,i-1}\sigma_{i-1}u_{i-1,t} + \sigma_i u_{i,t}. \tag{A.29}$$

As a result, the VAR(p) in equation (A.27) admits the following triangular structure,

$$\begin{aligned}
y_{1,t} &= c_1 + \boldsymbol{a}_{1,\cdot}\boldsymbol{Z}_t + \sigma_1 u_{1t}, \\
y_{2,t} &= c_2 + \boldsymbol{a}_{2,\cdot}\boldsymbol{Z}_t + \gamma_{2,1}\sigma_1 u_{1,t} + \sigma_2 u_{2,t}, \\
&\vdots \\
y_{n,t} &= c_n + \boldsymbol{a}_{n,\cdot}\boldsymbol{Z}_t + \gamma_{n,1}\sigma_1 u_{1,t} + ... + \gamma_{n,n-1}\sigma_{n-1}u_{n-1,t} + \sigma_n u_{n,t},
\end{aligned} \tag{A.30}$$

where $\boldsymbol{a}_{i,\cdot} = [a_{i,1}, ..., a_{i,np}]$ denotes the vector of coefficients in the i-th VAR equation, and $\boldsymbol{Z}_t = \left[\boldsymbol{y}'_{t-1}, ..., \boldsymbol{y}'_{t-p}\right]'$. As noted by Carriero et al. (2017), the re-parametrization of the VAR(p) in (A.30) allows for estimation of the system recursively, equation-by-equation.[37] For example, consider the generic equation i, which we rewrite as

$$y_{i,t} = c_i + \boldsymbol{a}_{i,\cdot}\boldsymbol{Z}_t + \gamma_{i,1}\sigma_1 u_{1,t} + ... + \gamma_{i,i-1}\sigma_{i-1}u_{i-1,t} + \sigma_i u_{i,t}, \tag{A.31}$$

Provided that all previous $i-1$ equations have been already estimated, all terms on the right hand side of (A.31) involving the previous equation error terms can be replaced by their estimated counterparts. As a result, the full posterior for the VAR parameters $\left\{\boldsymbol{c}, \boldsymbol{a}, \boldsymbol{\Gamma}^{-1}, \boldsymbol{\Sigma}\right\}$ can now be obtained recursively, one equation at a time.

[37] It is important to note that the triangularization in (A.30) produces the same posteriors for the coefficients that would be obtained by drawing the coefficients of all the equations simultaneously, and it does so regardless of the ordering in which the variables are entered in the VAR. However, it is worth to keep in mind that models where the priors are affected by the ordering will of course have posteriors which are also affected by such ordering. For example, if one were to elicit priors for $\boldsymbol{\Gamma}^{-1}$ and $\boldsymbol{\Sigma}$ separately, the implied prior for $\boldsymbol{\Omega}$ will change when the ordering of the equations in the VAR changes. As a result, different orderings of the variables in the VAR will lead to different prior specifications for $\boldsymbol{\Omega}$ and potentially different joint posteriors of the BVAR parameters $\{\boldsymbol{c}, \boldsymbol{a}, \boldsymbol{\Omega}\}$. As noted by Primiceri (2005), this problem will likely be less severe in the case as it is here in which the elements of the covariance matrix in $\boldsymbol{\Gamma}^{-1}$ do not vary with time, because the likelihood will quickly dominate the prior as the sample size increases. On this point, see also the estimation algorithms of Smith and Kohn (2002) and George et al. (2008) and discussions therein.

Appendix B Data and transformations

Table B.1. List of series

No	*Tcode*†	*Medium*	*Large*	*X-large*	*FRED*	*Description*
1	5	X	X	X	RPI	Real Personal Income
2	5	X	X	X	W875RX1	RPI ex. Transfers
3	5	X	X	X	DPCERA3M086SBEA	Real PCE
4	5	X	X	X	CMRMTSPLx	Real M& T Sales
5	5	X	X	X	RETAILx	Retail and Food Services Sales
6	5		X	X	INDPRO	IP Index
7	5			X	IPFPNSS	IP: Final Products and Supplies
8	5			X	IPFINAL	IP: Final Products
9	5			X	IPCONGD	IP: Consumer Goods
10	5			X	IPDCONGD	IP: Durable Consumer Goods
11	5			X	IPNCONGD	IP: Nondurable Consumer Goods
12	5			X	IPBUSEQ	IP: Business Equipment
13	5			X	IPMAT	IP: Materials
14	5			X	IPDMAT	IP: Durable Materials
15	5			X	IPNMAT	IP: Nondurable Materials
16	5			X	IPMANSICS	IP: Manufacturing
17	5			X	IPB51222S	IP: Residential Utilities
18	5			X	IPFUELS	IP: Fuels
19	2			X	CUMFNS	Capacity Utilization: Manufacturing
20	2		X	X	HWI	Help-Wanted Index for US
21	2		X	X	HWIURATIO	Help Wanted to Unemployed ratio
22	5		X	X	CLF16OV	Civilian Labor Force
23	5			X	CE16OV	Civilian Employment
24	2	X	X	X	UNRATE	Civilian Unemployment Rate
25	2			X	UEMPMEAN	Average Duration of Unemployment
26	5			X	UEMPLT5	Civilians Unemployed $\leq$ 5 Weeks
27	5			X	UEMP5TO14	Civilians Unemployed 5-14 Weeks
28	5			X	UEMP15OV	Civilians Unemployed $>$ 15 Weeks
29	5			X	UEMP15T26	Civilians Unemployed 15-26 Weeks
30	5			X	UEMP27OV	Civilians Unemployed $>$ 27 Weeks
31	5			X	CLAIMSx	Initial Claims
32	5	X	X	X	PAYEMS	All Employees: Total nonfarm
33	5			X	USGOOD	All Employees: Goods-Producing
34	5			X	CES1021000001	All Employees: Mining and Logging
35	5			X	USCONS	All Employees: Construction
36	5			X	MANEMP	All Employees: Manufacturing
37	5			X	DMANEMP	All Employees: Durable goods
38	5			X	NDMANEMP	All Employees: Nondurable goods
39	5			X	SRVPRD	All Employees: Service Industries
40	5			X	USTPU	All Employees: TT&U
41	5			X	USWTRADE	All Employees: Wholesale Trade
42	5			X	USTRADE	All Employees: Retail Trade
43	5			X	USFIRE	All Employees: Financial Activities
44	5			X	USGOVT	All Employees: Government
45	5		X	X	CES0600000007	Hours: Goods-Producing
46	2			X	AWOTMAN	Overtime Hours: Manufacturing
47	5			X	AWHMAN	Hours: Manufacturing
48	5			X	HOUST	Starts: Total
49	5			X	HOUSTNE	Starts: Northeast
50	5			X	HOUSTMW	Starts: Midwest
51	5			X	HOUSTS	Starts: South
52	5			X	HOUSTW	Starts: West
53	5			X	AMDMNOx	Orders: Durable Goods
54	5			X	AMDMUOx	Unfilled Orders: Durable Goods
55	5			X	BUSINVx	Total Business Inventories
56	2			X	ISRATIOx	Inventories to Sales Ratio
57	5		X	X	M1SL	M1 Money Stock
58	5		X	X	M2SL	M2 Money Stock
59	5		X	X	M2REAL	Real M2 Money Stock
60	5	X	X	X	BUSLOANS	Commercial and Industrial Loans

Table B.1 (continued)

61	5			X	REALLN	Real Estate Loans
62	5	X	X	X	NONREVSL	Total Nonrevolving Credit
63	2	X	X	X	CONSPI	Credit to PI ratio
64	5		X	X	S&P 500	S&P 500
65	5		X	X	S&P: indust	S&P Industrial
66	2		X	X	S&P div yield	S&P Divident yield
67	5		X	X	S&P PE ratio	S&P Price/Earnings ratio
68	2	X	X	X	FEDFUNDS	Effective Federal Funds Rate
69	2	X	X	X	CP3M	3-Month AA Comm. Paper Rate
70	2		X	X	TB3MS	3-Month T-bill
71	2		X	X	TB6MS	6-Month T-bill
72	2		X	X	GS1	1-Year T-bond
73	2		X	X	GS5	5-Year T-bond
74	2	X	X	X	GS10	10-Year T-bond
75	2		X	X	AAA	Aaa Corporate Bond Yield
76	2		X	X	BAA	Baa Corporate Bond Yield
77	1			X	COMPAPFF	CP - FFR spread
78	1			X	TB3SMFFM	3 Mo. - FFR spread
79	1			X	TB6SMFFM	6 Mo. - FFR spread
80	1			X	T1YFFM	1 yr. - FFR spread
81	1			X	T5YFFM	5 yr. - FFR spread
82	1			X	T10YFFM	10 yr. - FFR spread
83	1			X	AAAFFM	Aaa - FFR spread
84	1			X	BAAFFM	Baa - FFR spread
85	5	X	X	X	EXSZUS	Switzerland / U.S. FX Rate
86	5	X	X	X	EXJPUS	Japan / U.S. FX Rate
87	5	X	X	X	EXUSUK	U.S. / U.K. FX Rate
88	5	X	X	X	EXCAUS	Canada / U.S. FX Rate
89	5			X	WPSFD49107	PPI: Final demand less energy
90	5			X	WPSFD49501	PPI: Personal cons
91	5			X	WPSID61	PPI: Processed goods
92	5			X	WPSID62	PPI: Unprocessed goods
93	5		X	X	OILPRICEx	Crude Oil Prices: WTI
94	5			X	PPICMM	PPI: Commodities
95	6	X	X	X	CPIAUCSL	CPI: All Items
96	5			X	CPIAPPSL	CPI: Apparel
97	5			X	CPITRNSL	CPI: Transportation
98	5			X	CPIMEDSL	CPI: Medical Care
99	5			X	CUSR0000SAC	CPI: Commodities
100	5			X	CUUR0000SAD	CPI: Durables
101	5			X	CUSR0000SAS	CPI: Services
102	5			X	CPIULFSL	CPI: All Items Less Food
103	5			X	CUUR0000SA0L2	CPI: All items less shelter
104	5			X	CUSR0000SA0L5	CPI: All items less medical care
105	5			X	PCEPI	PCE: Chain-type Price Index
106	5			X	DDURRG3M086SBEA	PCE: Durable goods
107	5			X	DNDGRG3M086SBEA	PCE: Nondurable goods
108	5			X	DSERRG3M086SBEA	PCE: Services
109	5			X	CES0600000008	Ave. Hourly Earnings: Goods
110	5			X	CES2000000008	Ave. Hourly Earnings: Construction
111	5			X	CES3000000008	Ave. Hourly Earnings: Manufacturing
112	5			X	MZMSL	MZM Money Stock
113	5			X	DTCOLNVHFNM	Consumer Motor Vehicle Loans
114	5			X	DTCTHFNM	Total Consumer Loans and Leases
115	5		X	X	INVEST	Securities in Bank Credit
116	5	X	X	X	GDP	Real Gross Domestic Product
117	5			X	PCDG	PCE: Durable Goods
118	5			X	PCESV	PCE: Services
119	5			X	PCND	PCE: Nondurable Goods
120	5			X	FPI	Fixed Private Investment
121	5			X	PRFI	Private Residential Fixed Investment
122	5			X	GCEC1	Government Cons Expenditures Gross Inv
123	6	X	X	X	GDPDEFL	GDP deflator
124	5			X	PCEDEFL	PCE deflator

† Transformation code: 1 - levels; 2 - first differences; 5 - first differences of logarithms; 6 - second differences of logarithms